ARTIFICIAL NEURAL NETWORK BASED PREDICTION AND COOLING ENRGY OPTIMIZATION OF DATA CENTERS

A Dissertation
Presented to
The Academic Faculty

by

Jayati Athavale

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mechanical Engineering

Georgia Institute of Technology
May 2019

ARTIFICIAL NEURAL NETWORK BASED PREDICTION AND COOLING ENERGY OPTIMIZATION OF DATA CENTERS

Approved by:

Dr. Yogendra Joshi, Co-Advisor
The George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. Godfried Augenbroe
School of Architecture
Georgia Institute of Technology

Dr. Minami Yoda, Co-Advisor
The George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. Ada Gavrilovska
School of Computer Science
Georgia Institute of Technology

Dr. Satish Kumar
The George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Date Approved:

To Aai & Baba

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude towards my co-advisors Dr. Yogendra Joshi and Dr. Minami Yoda for giving me this opportunity of completing my dissertation under their guidance. I have truly appreciated their advice and guidance in matters regarding my thesis and in making other personal decisions as well. They have been an inspiration in a multitude of ways. I would also like to thank Dr. Satish Kumar, Dr. Godfried Augenbroe and Dr. Ada Gavrilovska for serving on my reading committee and providing their valuable insights.

I would like to thank the past and present members of my research group for their support, co-operation and a good working environment. The interactions and discussions I have had with them were of great learning value.

I am very grateful to my parents, my sister and my family for the unconditional love and unwavering support that I have received throughout my life. I would also like to thank my dear friend Rucha Jog for providing me with the love and comfort that only a long-standing friendship can provide. Her creativity and strength are an inspiration to me.

My time at Georgia Tech bestowed me with some wonderful and unique friendships. And these friends in turn ensured that I would cherish my time here forever. I would like to take this opportunity to thank all my friends for their support, care and most importantly for the innumerable fun-filled memories. You guys made everything easier (and more colorful)! I would be greatly amiss if at this point I do not mention Girish Kini and Subhrajit Chakraborty, both of whom have been a significant part of this journey.

Thank you for your advices, reprimands, reality-checks, patience, encouragement, criticism, belief and most importantly companionship. Girish will always remain a big part of the reason of why I was able to complete this journey.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVATIONS

A	Area for heat/mass transfer
b_{ANN}	Bias term in neural network
c_p	Specific heat
COP	Coefficient of Performance
F	Tile Porosity
k	Covariance function
K	Kernel Function for SVR and GPR Model
K_{coeff}	Pressure Coefficient
$\dot{m}$	Mass flow rate
n	Number of neurons in a layer
N	Blower/Fan Speed
NTU	Number of Transfer Units
p	Pressure
P	Power
$\dot{Q}$	Heat load
R	Set of real numbers
Re	Reynolds Number
T	Temperature
$\dot{V}$	Volume flow rate
w	Weight coefficients

x	Neuron in input/hidden layer
y	Neuron in output layer

Greek Symbols

α	Lagrangian Multipliers
ε	Error in SVR Model
ε_{HEx}	Heat exchanger effectiveness
ρ	Density
ψ	POD modes
θ	Hyperparameters for GPR Model

Subscripts/Superscripts

a	air
actual	actual value of property
amb	ambient
ANN	pertaining to Artificial Neural Network Model
chiller	chiller
con	condenser
CRAC	property of computer room air conditioner
eva	evaporator
GPR	pertaining to Gaussian Process Regression Model
POD	pertaining to Proper Orthogonal Decomposition Model
predicted	predicted value of property

rack	property of server rack
rack inlet	property over rack inlet face
ret	return
Room	property of data center room
sup	supply
SVR	pertaining to Support Vector Regression Model
tile	property of perforated floor tile
w	water

List of Abbreviations

Abbreviations

ANN	Artificial Neural Network
CFD/HT	Computational Fluid Dynamics/Heat Transfer
CRAC	Computer Room Air Conditioner
CUE	Carbon Usage Effectiveness
DDM	Data Driven Model
ERE	Energy Reuse Effectiveness
GA	Genetic Algorithm
GPR	Gaussian Process Regression
IT	Information Technology
LMA	Levenberg–Marquardt Algorithm
PDU	Power Distribution Unit
POD	Proper Orthogonal Decomposition

PUE	Power Usage Effectiveness
SVR	Support Vector Regression
TEP	Total Escaped Power
WUE	Water Usage Effectiveness

SUMMARY

Thermal management of data centers remains a challenge because of their ever-increasing power densities and decreasing server footprints. Current lack of dynamic control over global provisioning and local distribution of cooling resources often result in wasteful overcooling. These trends motivate this thesis research, which focuses on the development of a reliable and energy-efficient framework for allocating cooling resources to meet thermal management requirements, while minimizing energy consumption and adverse environmental impacts.

A key component of energy-efficient thermal management is real-time accurate prediction of temperature distribution in data centers. This first section of this dissertation focuses on development and comparison of four Data Driven Modeling (DDM) methods, namely Artificial Neural Networks (ANN), Support Vector Regression (SVR), Gaussian Process Regression (GPR) and Proper Orthogonal Decomposition (POD). These DDM methods were trained on datasets generated from offline Computational Fluid Dynamics/Heat Transfer (CFD/HT) simulations for real-time prediction of temperature and airflow distributions in a data center. Using CFD simulation results to train DDMs transfers computational complexity from model execution (in CFD) to model setup and development. To generate the training data, a physics-based and experimentally validated room-level CFD/HT model was developed using the commercial software Future Facilities 6Sigma Room.

Another key component of the overall framework is a model to estimate the cooling power consumed by a data center. This research developed a model based on thermodynamic analyses of data center cooling equipment, as described here.

Finally, development and implementation of a Genetic Algorithm (GA) based optimization framework in a data center lab is presented. The optimization framework employs an ANN based model to predict rack inlet air temperatures and a thermodynamic model to optimize cooling energy consumption. Results from a test run of 7.5 hours in the Data Center Laboratory indicate that implementing this optimization framework for dynamic provisioning of cooling resources reduces cooling power consumption by 20% compared with baseline operation without this optimization.

CHAPTER 1. INTRODUCTION

Data centers are large facilities that host computing and networking equipment for the purposes of collecting, storing, processing, distributing and accessing large volumes of data. Most enterprise operations depend on data centers and as such there were more than 8.5 million data centers worldwide in 2017 [1]. Typically, data centers have three basic components: 1) Information Technology (IT) Equipment, *i.e.*, the functional unit of data center operation, 2) Cooling Infrastructure, which cools the IT equipment and prevents thermal runaway, and lastly 3) Power Distribution Unit (PDU), which provides power to both the IT equipment and Cooling Infrastructure.

Figure 1 shows a side view of a data center room employing the most common cooling configuration for air-cooled data centers, namely underfloor supply and overhead return. The IT equipment is stored in racks arranged in a configuration with alternating hot and cold aisles, while the Computer Room Air Conditioning (CRAC) units are sited around the periphery of the room. The CRAC units pump cold air into the under-floor plenum, which enters the cold aisle through perforated tiles. This cold air is then drawn into the racks by server fans and exhausted to the hot aisle after passing over, and being heated by, the IT equipment. Finally, the heated air returns to the CRAC unit via the overhead return plenum, and the CRAC unit rejects this heat to cold water supplied by a chiller via an air-water heat exchanger. The ultimate rejection to the ambient is usually accomplished via air-cooled chillers or cooling towers.

Reference: J. Athavale, Y. Joshi, and M. Yoda, "Thermal Modeling of Data Centers," in *Advances in Heat Transfer, Volume 50,* (2018)

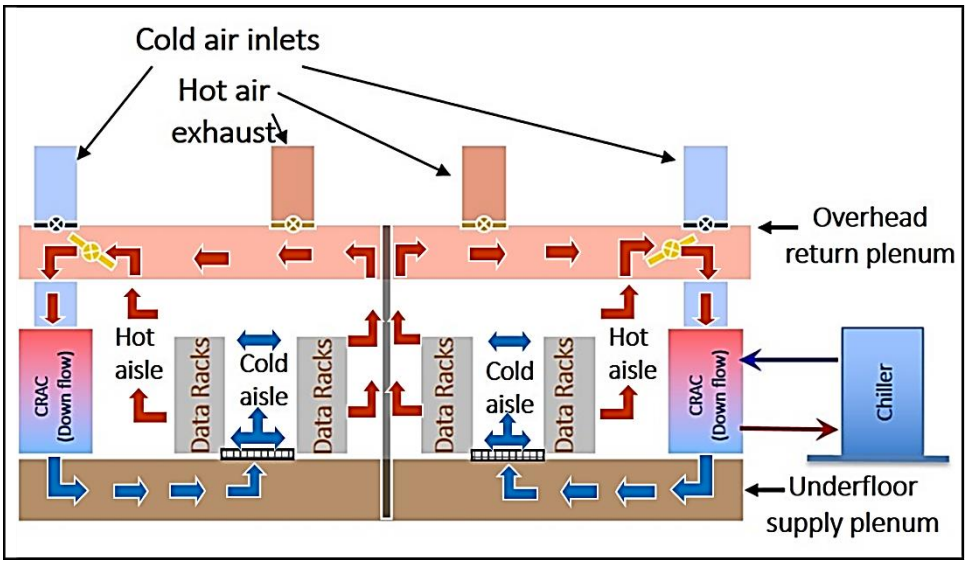


Figure 1- Common Data Center Cooling Scheme

1.1 Thermal Guidelines for Data Centers

The American Society of Heating Refrigeration and Air-Conditioning Engineers (ASHRAE) has developed and maintained environmental control guidelines for data centers. In 2004, ASHRAE TC 9.9 was compiled to provide a set of guidelines for air entering computing equipment [2]. This first version outlined conservative protocols and temperature thresholds (20 °C–25 °C), with the primary concern to ensure reliable operation of IT equipment; energy costs were a secondary concern for these guidelines. More recently in 2011, in view of growing concerns regarding cooling energy consumption of data centers, these guidelines have evolved to support energy reduction practices (higher temperature set-points) and technologies (free cooling and economization).

Data centers are divided into six classes: A1-A4, B and C based on equipment type, overall reliability needs and level of control as dictated by equipment and business specifications. Most data centers, including enterprise and volume servers and storage devices, as well as personal computers and workstations that require very stable

temperatures, fall in the A1 or A2 categories. Figure 2 and Table 1 provide details about these classifications and associated guidelines.

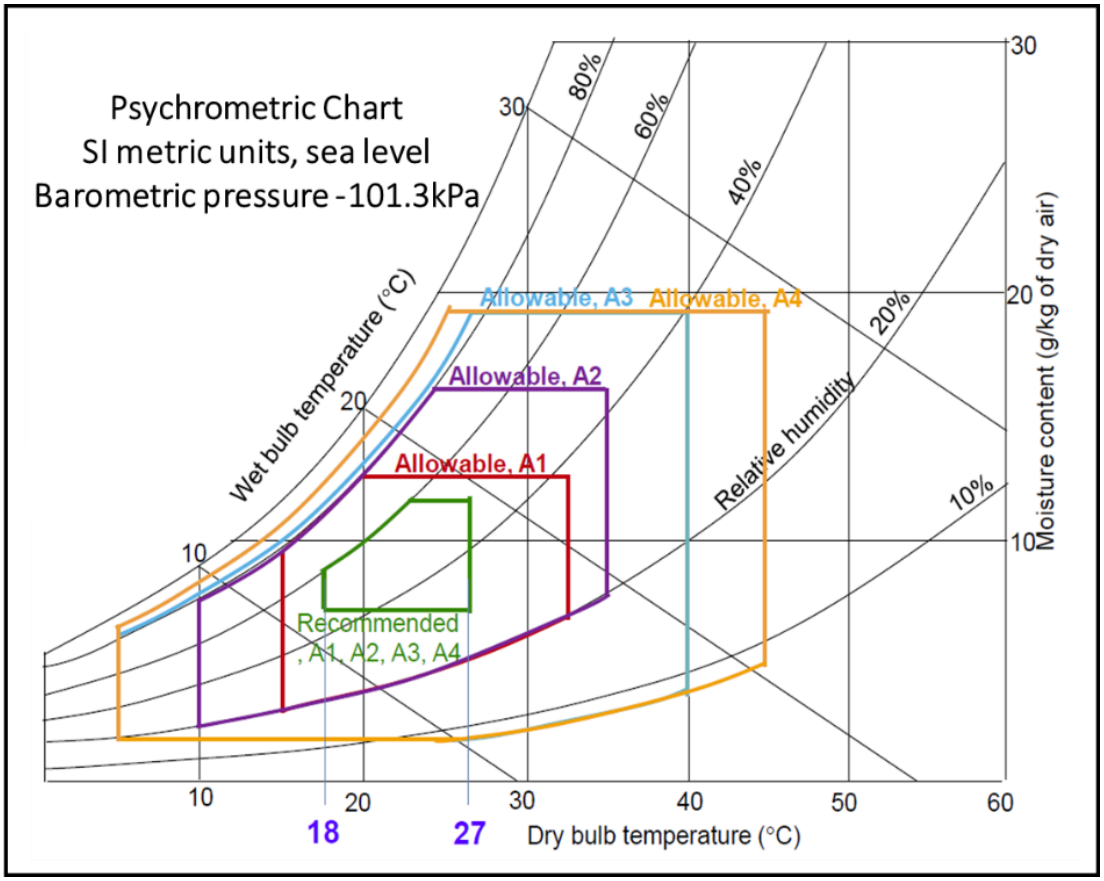


Figure 2 - ASHRAE Thermal Guidelines for Data Centers

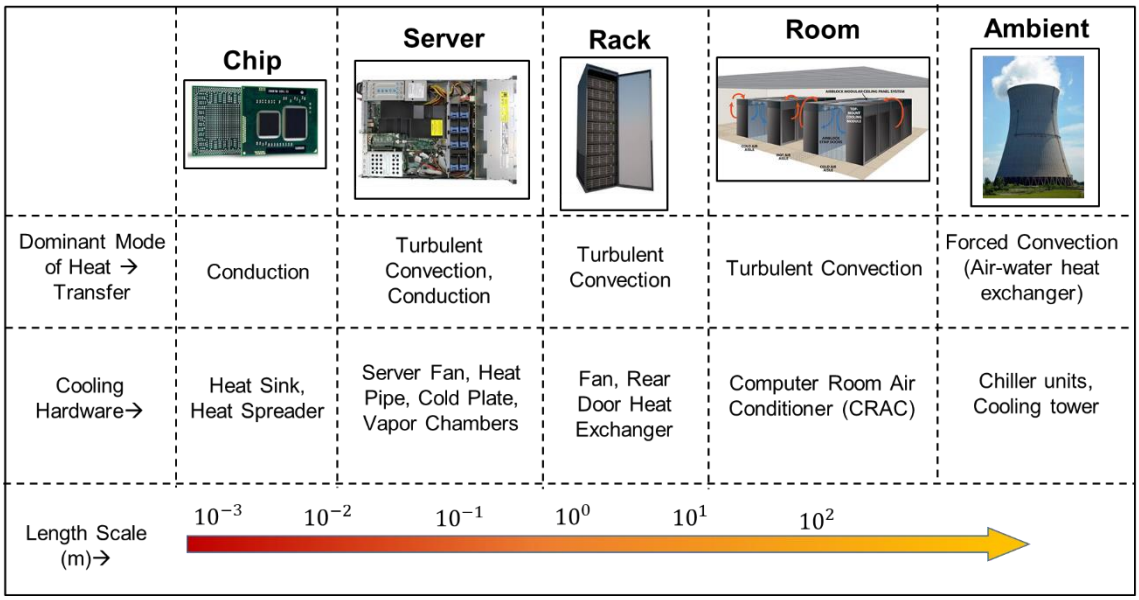
In Fig. 2, the recommended envelope describes the limits for most reliable operation of the IT equipment while still achieving reasonable energy efficiency for a data center. The allowable envelope represents the maximum limits under which IT equipment can be operated for short periods while maintaining functionality.

Table 1 - ASHRAE Thermal Guidelines for Data Centers

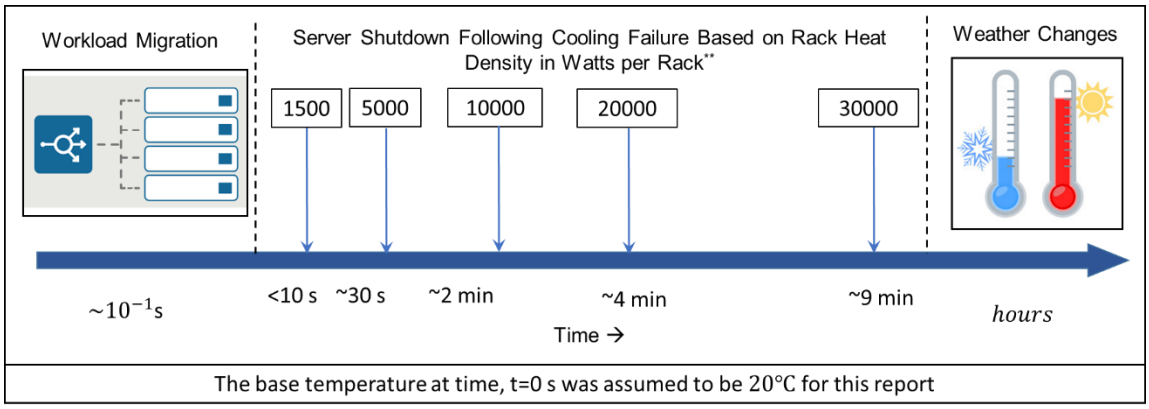
Class	Equipment Environmental Specifications for Air Cooling						
	Product Operations					Product Power Off	
	Dry-Bulb Temperature (°C)	Humidity Range, Non-Condensing	Maximum Dew Point (°C)	Maximum Elevation (m)	Maximum Temperature Change in an Hour (°C)	Dry-Bulb Temperature (°C)	Relative Humidity (%)
Recommended (Suitable for all 4 classes)							
A1 to A4	18 to 27	9°C DP to 15°C DP and 60% RH					
Allowable							
A1	15 to 32	12°C DP & 8% RH to 17°C DP and 80% RH	17	3050	20	5 to 45	8 to 80
A2	10 to 35	12°C DP & 8% RH to 17°C DP and 80% RH	21	3050	20	5 to 45	8 to 80
A3	5 to 40	12°C DP & 8% RH to 17°C DP and 80% RH	24	3050	20	5 to 45	8 to 80
A4	5 to 45	12°C DP & 8% RH to 17°C DP and 80% RH	24	3050	20	5 to 45	8 to 80
B	5 to 35	8% to 28°C DP and 80% RH	28	3050	NA	5 to 45	8 to 80
Class	5 to 40	8% to 28°C DP and 80% RH	28	3050	NA	5 to 45	8 to 80

1.2 Heat Transfer in Data Centers

Heat transfer in data centers occurs over a wide range of spatial and temporal scales (Fig. 3(a) and 3(b) [3]), making accurate design of efficient cooling schemes challenging. The chip and server are the components providing the core functionality of a data center, while the rack and data center room are part of the supporting infrastructure (mechanical, electrical and cooling) that ensure robust and reliable operation. Heat is primarily generated at the chip level, then dissipated to the air in the data center. The cooling hardware is primarily controlled at the data center, or room, level. Further adjustments are possible at local levels to achieve sufficient cooling at the server and chip levels. The cooling demands of the IT equipment in a data center are both spatially and temporally non-uniform due to random workload distributions and the time-varying nature of the workload itself. In order to meet the aforementioned thermal guidelines and ensure reliable operation of the IT equipment, dynamic actuation and control of cooling infrastructure is essential. Furthermore, implementation of real-time control makes it possible to use higher temperature set-points without the risk of server over-heating, and can therefore reduce cooling energy consumption. A survey in 2013 indicated that 90% of data centers operate at temperatures under the set-point of 24 °C, suggesting that they are overcooled and wasting energy [4].



(a)



(b)

Figure 3(a) - Varying Spatial Scales and Modes of Heat Transfer in Data Centers (b) - Varying Temporal Scales and Evolution of Typical Transient Scenarios [2]

1.3 Energy Scenario and Need for Dynamic Control

Data centers are mission-critical facilities. As such, the primary design objective for the cooling infrastructure is to provide an acceptable thermal environment at all times, and minimize or, if possible, eliminate cooling failure-linked downtime. A national Survey on Data Center Outages conducted by the Ponemon Institute in 2013 indicated that 95% of the data centers surveyed had experienced outages; the average cost to the business of these

outages was $8000/min [5]. Root-cause analysis indicated that heat-related issues/computer room air conditioner (CRAC) failure were among the top seven causes of these outages.

In 2014, data centers in the U.S. consumed an estimated 70 billion kWh, representing about 2% of total U.S. electricity consumption, and this consumption is estimated to increase to 75 billion kWh by 2020 [1]. Benchmarking studies [4] reveal that cooling infrastructure accounts for 30-50% of the total power consumed in a data center; in the worst-case scenarios, facility- side power consumption exceeds that for the IT equipment (Fig. 4).

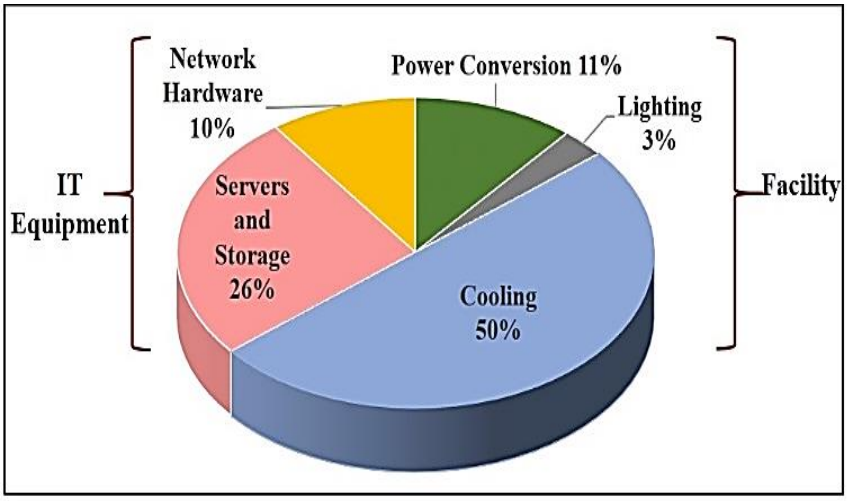


Figure 4 - Energy Consumption Breakdown in Data Centers [3]

The energy consumed by data centers for cooling purposes can be broadly divided into two parts: 1) energy consumed by chiller, which runs on a refrigeration cycle, and 2) energy consumed by fluid (air, water or refrigerant) transport components like CRAC blowers, refrigerant and chilled water pumps and server fans. Figure 5 shows the cooling energy breakdown for a typical air-cooled data center facility [6]. As can be seen, chiller

energy use is the biggest fraction of the total cooling energy used and is also reported to have the lowest efficiency, followed by CRAC energy use [6].

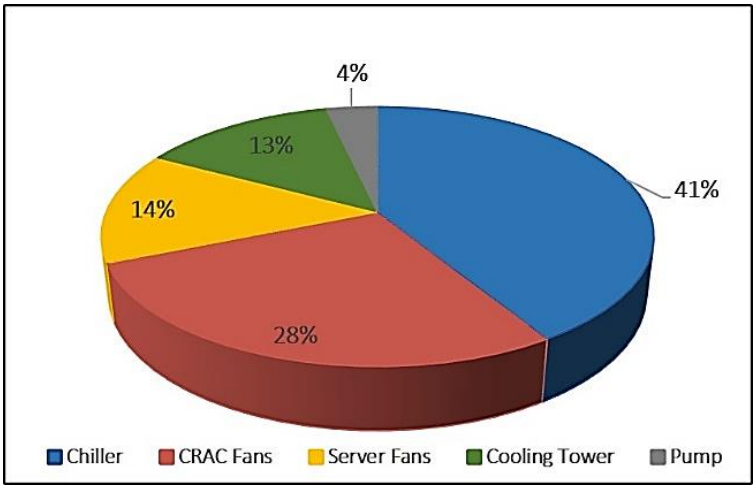


Figure 5 - Cooling Energy Breakdown

1.3.1 Metrics for Energy Efficiency in Data Centers

The Power Usage Effectiveness (PUE) can be used to characterize the energy efficiency of data centers. This metric represents the ratio of the energy consumed by infrastructure (cooling equipment, uninterrupted power supplies, lighting etc.,) to that consumed by IT equipment:

$$PUE = \frac{\text{Total Facility Energy}}{\text{IT Equipment Energy}} \tag{1.1}$$

A PUE value near one indicates energy-efficient operation for a data center. Figure 6 shows PUE values reported for 1100 data centers in a survey conducted by the Uptime Institute [4].

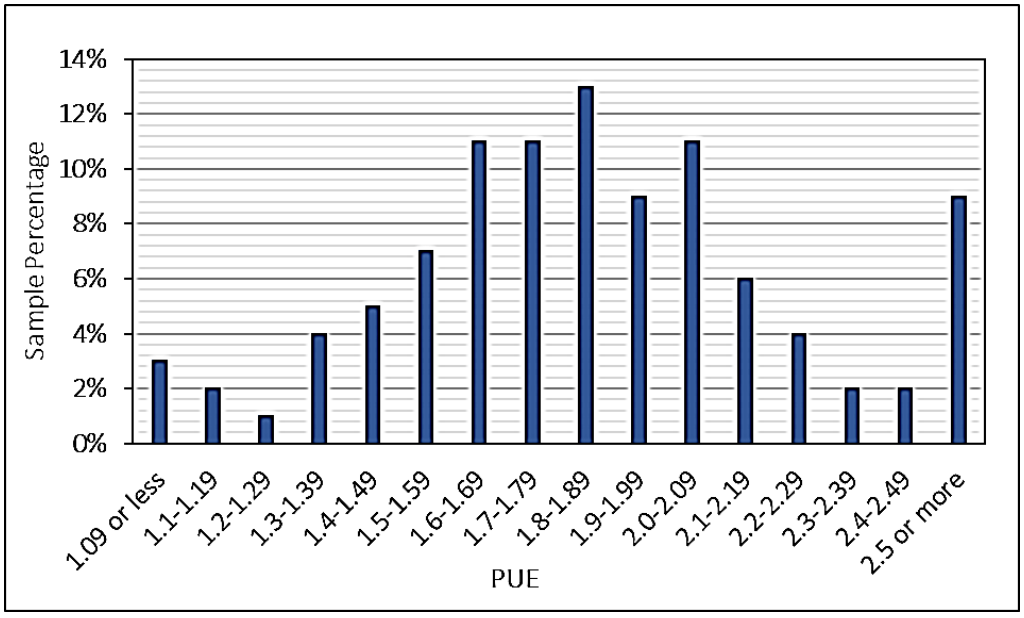


Figure 6 - PUE Values Reported in Uptime Institute Data

With concerns about the carbon footprint and greenhouse emission of data centers and their effect on the environment increasing, characterizing the "greenness" of data centers has become important. A number of metrics to evaluate sustainability of data centers have been defined and are presented below in Table 2 [7-9].

Table 2 - Sustainability Metrics for Data Centers

Metric	Definition	Unit	Objective	Ideal Value
Carbon Usage Effectiveness (CUE) [4]	$\frac{\text{Total } CO_2 \text{ emissions caused by total data center energy usage}}{\text{IT Equipment Energy}}$	$(KgCO_2 / kWh)$	Minimize	0.0
Energy Reuse Effectiveness (ERE) [5]	$\frac{\text{Total Energy-Reuse Energy}}{\text{IT Equipment Energy}}$	(-)	Minimize	0.0
Water Usage Effectiveness (WUE) [6]	$\frac{\text{Annual Site Water Usage}}{\text{IT Equipment Energy}}$	$(Litres / kWh)$	Minimize	0.0
Water Usage Effectiveness at Source (WUE_{source}) [6]	$\frac{\text{Annual Source Energy Water Usage} + \text{Annual Site Water Usage}}{\text{IT Equipment Energy}}$	$(Litres / kWh)$	Minimize	0.0

Recent studies have shown that despite the wide range of PUE values across data centers, the average value reported for 2014 was 1.8, only a modest improvement over the average of 2.0 reported in 2007 [10]. A PUE of 1.8 indicates that energy consumed by supporting infrastructure (*e.g.* for cooling) is 80% of that consumed by the IT equipment. One of the reasons for the slow rate of improvement in PUE is the "life cycle mismatch" between the IT equipment and cooling equipment, where the turnover in a data center's infrastructure is much slower than that of the servers. Although current data center infrastructure and cooling equipment are built and designed to allow for flexibility and retrofitting, reliability concerns and lack of high-fidelity control frameworks lead to using static conservative cooling set-points based on maximum IT capacity. Most data centers at present are hence significantly overcooled.

These trends, when coupled with increasing energy costs and associated environmental impacts, have shifted the emphasis of research in this area from thermal management alone to energy optimization and thermal management schemes that can reduce energy consumption without compromising server reliability. A Lawrence Berkeley National Laboratory study, for example, estimates that adoption of energy-efficient operations by data centers could lead to annual savings of up to 33 billion kWh by 2020, or a 45% reduction in electricity demand, further motivating this research [10].

1.3.2 Typical Components of Dynamic Control Framework in Data Centers

A framework for a dynamic control system designed to meet temperature thresholds in data centers consists of three major components: a data collection platform (including

additionally installed as well as onboard server sensors), a model for temperature prediction and a trained control algorithm, as shown in Fig. 7.

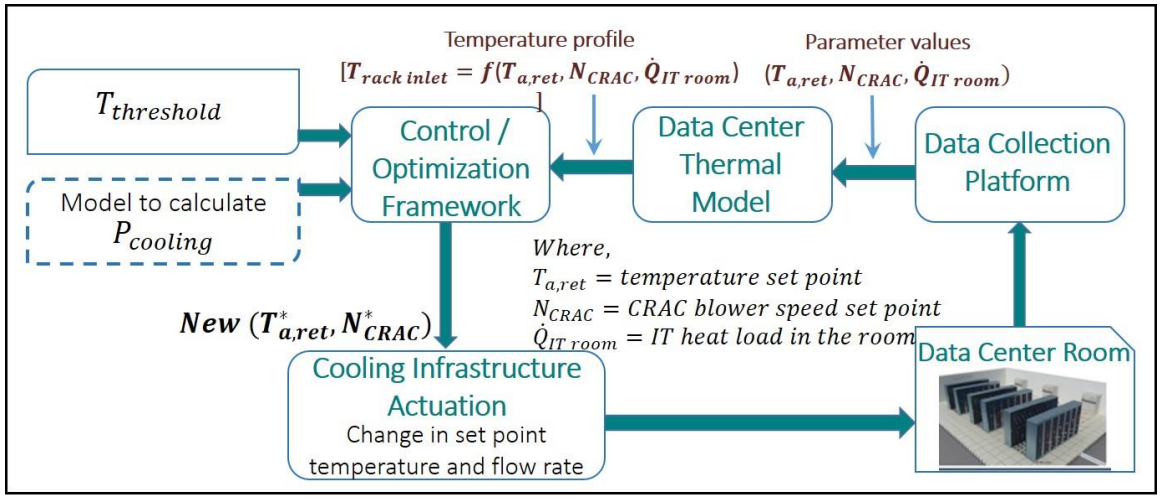


Figure 7 - Dynamic Control Framework for Cooling Control in Data Center

The data collection platform is used to aggregate information such as cooling set-points, rack inlet temperature and CPU utilization, for a given data center operating state. The data center model then uses the input data to predict temperature and/or airflow distributions for a particular operating state. Finally, the control algorithm determines the control action for cooling infrastructure based on these predictions and the desired temperature distribution. In addition to these three components, a model to estimate cooling energy consumption is required when energy optimization is also of interest (block in dashed line).

In air-cooled data centers, accurate predictions of the temperature field are a challenge because of the airflows in these complex geometries, as well as the coupling of flow with the heat transfer within the servers. Yet such predictions are required to provision and efficiently distribute the cold air to ensure that each and every server functions within

its thermal thresholds, while at the same time minimizing the associated energy consumption. The following section (section 1.4) provides a high-level discussion of different data center thermal and flow modeling approaches and their suitability for a dynamic optimization framework. Sections 1.5 and 1.6 review previous studies on dynamic control and optimization.

1.4 Modeling Approaches for Data Centers

A number of different approaches have been considered for modeling the airflow and thermal transport in data centers. These techniques can be broadly classified into three categories: 1) Simplified/ lumped-capacitance modeling approaches, 2) CFD/Heat Transfer (HT)-based numerical modeling approaches, and 3) data driven reduced-order modeling approaches. Figure 8 compares these three approaches in terms of their accuracy, information density and execution time.

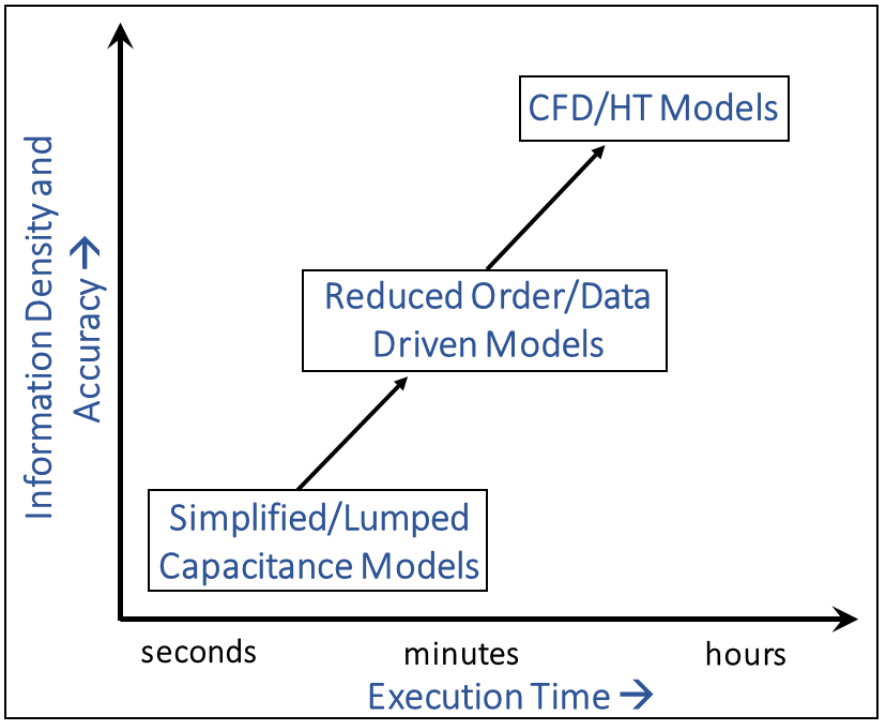


Figure 8 - Comparison of Modeling Frameworks

1.4.1 Simplified/Lumped Capacitance Modeling

Thermodynamics-based simplified models have shorter execution times compared with CFD/HT-based methods and are therefore useful for parametric studies and rapid prediction of temperature and airflow distribution. At the server level, Resistor-Capacitor (RC) thermal models were developed in [11, 12] to simulate the transient behavior of server components. Thermodynamics-based models, employing either energy [13, 14] or exergy [15] analyses of data center facilities have been developed for the most part assuming a uniform temperature over each spatial zone. Ref [16] explores a flow-network modeling technique for characterizing airflow in data centers as a function of the pressure difference across perforated tiles. Typically, the lumped-capacitance/simplified models lose many of the degrees of freedom characteristic of convective heat transfer due to assumptions inherent to the modeling framework. Hence, while capable of rapid temperature

predictions, reduced-order models in most cases are unable to provide the finer-scale predictions required for optimization frameworks to ensure reliable operation of each server.

1.4.2 Computational Fluid Dynamics Modeling

CFD/HT has been extensively used to model thermal transport and airflow in data centers. Table 3 summarizes many of the modeling efforts employing such framework.

Table 3 - Representative CFD/HT Studies for Data Centers

Author	Year	Scope	Ref
Kang et. al	2000	Plenum design	[17]
Schmidt et al.	2001,2002 ,2007	Plenum design, Tile design, Experimental Benchmarking, Design decision, Experimental Validation	[18-20]
Karki et. al	2003	Plenum design, flow through perforated tiles	[21]
Rambo et al.	2003	Multi-scale room level modeling	[22]
Bhopte et. al	2006	Effect of model complexity on predictions, Effect of underfloor blockage	[23, 24]
Iyengar et al., Cruz et al., Abdelmaksoud et. al	2007,2009 ,2010	Modeling and experimental validation of a small test cell, Comparison of turbulence models for data center modeling	[25-27]
Radmehr et. al, Erden et. al, Arghode et. al	2007,2013 ,2014	Rack level modeling, Rapid modeling (modified body force model) of airflow through tiles	[28-30]
VanGilder et. al	2007	Partially decoupled aisle based CFD modeling	[31]
Nelson et. al, Choi et. al, Pardey et.al	2007,2007 ,2015	Server level modeling	[32-34]
VanGilder et. al	2008	Effect of gird size on data center modeling	[35]
Patankar.,S., Wibron., E.	2009, 2015	Room level modeling	[36, 37]
Ibrahim et. al, Gondipalli et, al	2010,2010	Transient modeling, Effect of time-dependent boundary conditions	[38, 39]
Almoli. A	2013	Hot and cold aisle containment	[40]
Cruz et. al	2015	Coupled inviscid-viscous solution method	[41]

Even after significant simplification of various components, CFD/HT modeling involves discretizing transient, non-linear (convection terms), three-dimensional, second-order (diffusion term) coupled partial differential equations (namely, the three components of the momentum equation and the energy equation) and the three-dimensional continuity equation [42]. These equations are solved iteratively, so the required computational resources scale with the cube of the product of the number of spatial grid points and time steps. For most data centers modular configurations may repeat, the number of spatial grid points is of the order of millions, while the number of time steps is typically of the order of thousands, depending on the type of transient problem studied. Hence a CFD/HT modeling framework, while being accurate and offering high information density, is too time-intensive for a dynamic optimization framework. A more detailed review of CFD/HT based modeling efforts is presented in [43-45].

1.4.3 Reduced Order/Compact Modeling

Reduced-order models offer an acceptable tradeoff between the two approaches discussed previously in that that they can preserve the granularity and accuracy of CFD/HT type frameworks while achieving execution speeds comparable to simplified models, as indicated in Fig. 8. Compact models typically employ statistical or regression-based tools, in combination with physics-based modeling and/or experimental data for system identification and characterization, to predict system behavior. Development of compact models for data centers can involve geometry-based (e.g. fixed equipment layout), configuration-based (e.g. fixed recirculation pattern), or physics-based (e.g. uniform heat generation in servers) assumptions to achieve the desired reduction in modeling effort. These models are suitable for dynamic optimization frameworks, and can also be used to

improve the parametric granularity of data set, obtained either by experiments or simulations. Such compact models can be classified in four distinct categories, based upon the problem formulation, required data inputs and solution methodology:

1. Physics Based Models
2. Heuristic Models
3. Data Driven Models
4. Hybrid/Combination type Models

These classifications do not distinguish *per se* based on the physical domain of the model, so each category comprises of both component- or room-level models. A detailed review of data driven compact models is presented in Chapter 3 due to their relevance to this thesis. However, a high level review of studies employing alternate compact modeling approaches (listed above) in data centers that can provide predictions in (nearly) real-time for dynamic control of thermal management hardware is included in table 4 with more detailed discussion in [46].

Table 4 - Overview of Reduced Order Modeling of Data Centers
(*- indicates experimental validation is included) [46]

Category	Author/ References	Year	Scope		Geometry/Physical Domain	Runtime	Accuracy
Physics Based Reduced Order Modeling			**Flow Modeling**	**Thermal Modeling**			
	Cruz et al.,	2009 and 2013	Inviscid CFD Model	Energy Equation	$84m^2$ area with single rack, one CRAC and 3 tiles	less than 100s	RMS error in temperature prediction ~2°C
		2015	Coupled Inviscid-Viscous Solution Method	Energy Equation	$84m^2$ area with single rack, one CRAC and 3 tiles	~700-1200 s depending on turbulence model employed	RMS error in temperature prediction ~3°C
	Toulouse et al.,*	2009 and 2010	Potential Flow based model	Convective transport model	21 ft. x26 ft. room with 17 racks	~12 s	Average error in temperature prediction of 5.0°C with high local maximum error of 21.7°C
	Lettieri et al.,	2013	Potential flow based model with Rankine Vortex superposition to account for buoyancy	Convective transport model	21 ft. x26 ft. room with 17 racks	~23 s	Average error in temperature prediction of 2.4°C with high local maximum error of 11.2°C
	Lopez and Hamnann*	2010,2011 and 2013	Potential Flow Based Model	Convection-Diffusion Equation Model	Space $238m^3$ in volume with 15 racks and 2 CRAC units	~30s	Absolute error for 90% points was in range [-2,+2]°C

	Shrivastava et al.,	2006 and 2007	Partially decoupled aisle method to predict rack CI		Single cluster with 12 racks and 4 local coolers	10-30 s	RMSE 5.5% with maximum error 28.8% in CI prediction
	Van Gilder et al.,	2006					
	Song et al.,	2013	Zonal Modeling Approach with mass and energy balance equations and ideal gas law assumption		1 m^3 volume with single rack divided into 27 zones	2s	Average relative error for temperature prediction 10%
	Song et al.,	2014	Zonal modeling with determination of inter-zonal boundary conditions using full scale CFD		11m x 8.4 m room area with 4 rows of 15 racks rack	50% enhancement as compared to full scale CFD (3.6 hours)	Average relative error of 9.7% with maximum localized error over 20%
Heuristic Methods	Tang et al.,	2006	State Space model, cross-interference pattern for recirculation characteristics obtained using CFD simulations			less than 1 min	Average prediction error in temperature~0.4°C
	Jonas et al.,	2012	State Space model, cross-interference pattern for recirculation characteristics obtained using CFD simulations		Two rows of 4 racks each (total 288 servers)	(-)	Average discrepancy in temperature prediction 0.8°C
	Li et al.,*	2011	(-)	Time series forecasting for temperature evolution	Three rack section in larger data center room	~real time for 5 minute look ahead window	
	Chen et al.,*	2012	(-)	Time Series Forecasting for temperature evolution	1) Single rack test bed, 15 servers , 30 ft^2 room size 2) Five racks,229 servers in production data center	~real-time for 4 minute look ahead window	Average error in rack inlet temperature prediction <2°C for 90% instances

Data Driven Methods	Samadiani et al.,	2010	POD model for temperature prediction for different CRAC velocities and heat load; training data gathered via CFD simulations	Simulated room 23.2m^2 in area with 32 racks and 4 CRAC units	runs 150 times faster than full scale CFD for the same configuration	Average prediction error for temperature ~1.24^{o}C	
	Samadiani et al.,*	2012	POD model for temperature prediction for different CRAC speeds; training data gathered via experimental measurements	102.2m^2 test facility with 15 racks with two CRAC units	~2s	Average prediction error for temperature~0.68^{o}C; Maximum localized error ~8^{o}C	
	Ghosh et al.,	2011	POD model for transient temperature prediction; training data gathered via CFD simulations	Data center room with eight racks and 1 CRAC unit	~4s		
	Ghosh et al.,	2014	POD model for transient temperature prediction for different IT loads; training data gathered via experimental measurements	6.4mx 8.5m test facility - only 1 rack and 1 CRAC unit operational during the study	~10s	Root mean square error for temperature prediction ~4.7%	
	Song et al.,	2013	POD model for transient flow prediction with varying CRAC operating conditions; training data gathered via CFD simulations	6.4mx 8.75m test facility - only 1 rack and 1 CRAC unit operational during the study	in order of seconds	Relative error of 2% in air velocity prediction	

	Song et al.,	2014	POD and NLPCA model for transient flow and temperature prediction with varying CRAC operating conditions; training data gathered via CFD simulations	6.4mx 8.75m test facility - only 1 rack and 1 CRAC unit operational during the study	(-)	Maximum relative error in temperature prediction -10% for POD based model and ~20% for NLPCA based model	
	Moore et al.,	2006	ANN based model for workload placement	Four rows of 7 racks each (total 1120 servers)	(-)	Average prediction error <1°C for 90% of instances	
	Gao, J.	2014	ANN based model for Power Usage Effectiveness (PUE) prediction and sensitivity analysis; training data obtained from monitored senor data		(-)	Relative error in PUE prediction 0.4%	
	Shrivastava et al.,	2007	ANN based model for Capture Index (CI) prediction for cooling clusters; training data obtained from PDA-CFD simulations	14ft long cluster of racks with in-row coolers	(-)	Relative error in CI prediction <10% for 96% of cases; Maximum error 27.4%	
	Song et al.,	2014	ANN model to parametrically study effect of plenum characteristics on tile flow rate and rack inlet temperature; training data obtained via CFD simulations	Room 11mx4.2m with one cold aisle and 15 racks	~real-time prediction	Relative error in temperature prediction 5%	

	Athavale et al.,	2018	ANN model for tile flow rate, rack inlet temperature and CRAC flow rates and temperature; training data obtained from experimentally validated CFD model	6.2m x 8.7m room with 12 racks and 1 CRAC unit	2s	Average error <0.6^{o}C for rack inlet temperature prediction; relative error of 0.7% for tile flow rate prediction	

1.5 Data Center Thermal Management

Table 5 reviews various studies of dynamic control for cooling infrastructure in data centers both on the component and room-level. Most of these studies control rack inlet temperature, as advised by the ASHRAE Technical Committee 9.9, which deals with mission critical facilities. It should be noted that all the studies mentioned have thermal management as their sole objective; any additional gains in energy efficiency are a secondary outcome. The following section describes room-level studies.

Table 5 - Overview of Dynamic Control Studies

	Ref.	Controlled Parameter	Manipulated Parameter	Control Methodology
Component Level Studies	[47-50]	Rack Inlet Temperature	Adaptive Vent Tile (AVT) Opening	MIMO Proportional Integral (PI) Controller
	[51]	Rack Inlet Temperature	Server Fan Speed	MIMO fan controller with convex optimization
	[52]	CPU temperature for single rack	CRAC Supply Temperature set point	Integral Control
	[53]	Rack Inlet Temperature	Active Tile Fan Speed	Manually tuned
Room Level Studies	[54-56]	Rack Inlet Temperature	CRAC Supply Temperature, CRAC Blower Speed	Cascaded PID Control
	[57, 58]	Rack Inlet Temperature, plenum pressure	Adaptive Vent Tiles , CRAC Blower Speed	Model Predictive Controller
	[59]	Mismatch between airflow supplied by CRAC unit and that required by IT equipment	CRAC blower speed	(-)
	[60]	On-board server inlet air temperature	CRAC Supply Temperature	(-)

Data center thermal management challenges have steadily increased over the last few years due to an increase in rack-level power densities, combined with largely static cooling set-points. To overcome these challenges Boucher et al. [61] conducted a study to test the viability of dynamic cooling control in data centers. Experiments were conducted by varying parameters such as the CRAC supply temperature, blower speed, and tile vent opening to determine their effects on rack inlet temperature and Supply Heat Index (SHI). Their results demonstrated that CRAC supply temperature and tile vent openings have a

predictable effect on rack inlet temperature and hence are ideal as control variables. Furthermore, the results also indicated that a combination of CRAC supply temperature and blower speed could be employed in energy optimization studies for data centers.

A conceptualized design for a "Smart Data Center", consisting of distributed sensing, flexible actuators and control policies for different components of data center was presented by Patel el al. [54, 55]. The study proposes implementing coordinated control at the chip, system and complete data center levels for globally efficient operation. In [56], Bash et al. explored dynamically and locally allocating cooling resources, as required by modulating the set-points for the CRAC units based on local data gathered from the corresponding regions of influence for the CRACs, thereby reducing energy consumption. Thermal models for the control system or "plant function" were identified by conducting basic system identification experiments to determine the regions of influence for the CRAC units, and a cascaded PID control for CRAC supply temperature and airflow rate was implemented. This control architecture could efficiently allocate cooling resources for individual sub-sections of the data center, and results for a test case showed a reduction of more than 25% in energy consumption compared to conventional operation (constant CRAC blower speed of 95% and return air temperature control). However, the architecture was found to be inefficient at smaller, i.e., individual rack or server scales, because that required controlling air delivery at these scales. Also, as no explicit thermal model is included in the framework, the control action is heuristic and based purely on tuning parameters for the PID controller. Hence, this framework may not be effective for all operational cases.

Wang et al. [57, 58] proposed a layered controller that coordinated local vent tile tuning and CRAC blower control to achieve thermal management in data centers. The test bed consisted of two rows of racks with a total of 17 racks and two CRAC units with a cold aisle populated with 20 Adaptive Vent Tiles (AVT). Control over rack inlet temperatures was exerted by actuating the AVT, while plenum pressure was maintained constant by manipulating CRAC blower speed using Variable Frequency Drives [57]. A model predictive controller was implemented for the AVT with the model parameters identified online while a Proportional Integral (PI) controller was implemented to tune CRAC blower speed. A limitation of this approach, as mentioned in regards to [49] as well, is that the rack inlet temperature at next instant is assumed to depend solely on its current value and vent tile opening (and therefore mass flow rate of air being supplied locally), neglecting the effect of CRAC settings and workloads. This translates to approximating non-linear correlations between rack inlet temperature and vent tile openings using linear Auto-Regressive Moving Average (ARMA) models, which would be valid only for certain regions of operation.

A mismatch between the airflow supplied by the CRAC unit and that required by the IT equipment results in undesirable airflow patterns like bypass, leakage and recirculation, decreasing the overall efficiency of the cooling configuration. Room-level control targeted at minimizing this mismatch was developed by Ahuja et al. [59] which involved dynamically matching the airflow supplied by the facility fans with aggregate airflow required by the servers. A linear relationship between volumetric airflow through a server and the server fan speeds was obtained by characterizing various servers using wind tunnel measurements. Additionally, the thermal sensor control for the CRAC unit

was moved from the return to supply side, bringing it closer to server inlet temperature, where the thermal management criterion is assessed. This enabled use of higher temperature set-points. The performance of the control framework was evaluated using simulations for a 1 MW facility. A reduction of 77% in the energy consumption of CRAC units is reported, as compared to conventional operation (no CRAC flow control, and return air temperature set-point for CRAC unit). It should be noted, however, that while going from return air temperature control to supply air temperature control for the CRAC units, the set-point temperature was kept the same (=21 °C), which would indicate that the data center room was being grossly over-cooled in the original design and hence the actual energy savings in another case may be much lower.

Zhang, et al. [60] investigated and compared room-level CRAC supply air temperature control based on feedback from sensors located at two different locations: 1) sensor located on the ceiling in middle of cold aisle; and 2) On-board inlet temperature sensors for the servers. The control strategies were tested using a CFD model of single cold aisle data center room with eight racks on each side of the cold aisle. Constructing a control system employing feedback from server inlet temperatures resulted in 3 °C increase in supply temperature set-point, which directly translates to a decrease in chiller energy consumption. Also, as the control action directly tracks the server inlet temperature, it is acceptable to decrease safety margins, further increasing the potential for energy savings. Energy calculations demonstrated a reduction in annual PUE from 1.31 to 1.21 between the two cases.

1.6 Data Center Thermal Management with Energy Optimization

Optimization studies that simultaneously consider thermal management and cooling energy optimization can be divided into two categories, Static Optimization and Dynamic Optimization, depending on whether the optimization framework runs in real-, or nearly real-time. Results from static optimization studies generally inform data center design decisions and involve optimizing data center layout, arrangement and number of aisle in a data center and number of racks in different aisles [62, 63]. The framework can also be implemented to predict the "ideal range" for some data center operation parameters (e.g., target temperature rise across servers and across CRAC units) with consideration of efficient operation of data centers and cooling infrastructure[64-68].

Dynamic optimization frameworks run in real time and dictate either changes to cooling set-points, IT load allocation and/or migration or both to ensure reliable operation of data centers within temperature thresholds and minimum possible energy consumption. These frameworks are also distinguished by their use of active control.

1.6.1 Static/Configuration based Optimization

Exergy-based static optimizations have been explored in [64-68] wherein the total exergy destruction in a data center room is assumed to be the sum of the exergy destruction in CRAC, racks and the air space. The CRAC unit is modeled as simple air-conditioning unit (open system) and corresponding exergy loss is estimated based on flow exergy for inlet and outlet streams of air and electricity consumed (work done) by the CRAC unit. Racks are modeled as single computed unit dissipating heat at a uniform temperature wherein the predominant exergy loss is attributed to the conversion of high-quality

electrical energy to low-quality thermal energy. Finally, exergy loss in the airspace is estimated by dividing the physical volume into a mesh of smaller volumes and calculating exergy loss for each cell. This approach assumes constant and fixed recirculation patterns in data center room, and fixed heat dissipation and exergy destruction in racks [64]. Studies to predict optimal CRAC supply temperature and flow rate for CRAC unit(s) that minimize exergy destruction for a single CRAC system [65] and multi-CRAC systems [66, 67] were conducted. The objective for optimization framework was to maximize the second-law efficiency for a data center room with respect to the temperature and flow set-points for the CRAC units for a given data center physical and IT load configuration. Experimental validations indicated that the model for prediction of room-level exergy destruction was accurate for low heat loads, but had errors as great as 25% at high heat loads. It should be noted that the framework inherently assumes that optimization of exergy loss in conjunction with appropriate provisioning of cooling resources will suffice for both thermal management as well as energy efficiency considerations and thus excludes explicit monitoring or modeling of rack inlet temperature for validation, which could potentially pose reliability concerns for data center operation.

Li et al. have considered design optimization for an enclosed data center cabinet using Multi-Objective Genetic Algorithms (MOGA) [69]. The two design objectives for the problem are: 1) minimizing the maximum chip temperature, and 2) maximizing the sum of total heat generation rates, are correlated and conflict with a constraint on maximum chip temperature. A POD-based reduced-order model of the data center cabinet is used to predict chip temperatures as required by the objective and constraint function. Appropriate values and bounds were obtained for the heat generation rate and chip temperature using

the optimization framework. It should be noted that this study is a rack-level design optimization, and only considers thermal management; minimization of cooling cost is not considered.

Optimization of room cluster layout based on parameters like rack capture index (CI) and room Total Escaped Power (TEP) was considered by Shrivastava et al. [62]. The cluster layouts employ local in-row cooling for removal of heat dissipated by the IT racks. CI, defined for each rack, is the fraction of air exhausted by the particular rack that is captured by local extracts (in-row coolers) while TEP assesses the overall performance of the entire cluster and is defined as fraction of heat dissipated by the cluster that is captured by in-row coolers. Lower values CI for individual racks in a cluster and subsequently high TEP is indicative of inefficient operation with increased recirculation and hotspots. An ANN model [70] relating cluster layout details (input) to CI and TEP values (predicted output) is developed and employed in conjunction with Genetic Algorithm(GA)-based optimization for generating and evaluating cluster layout alternatives for different objectives, including finding the best layout for fixed population of equipment, or finding the best location for additional heat load.

A method for design optimization for data center using Compromise Decision Support Problem (cDSP) framework has been presented in [71]. The conceptual basis of cDSP is to minimize the difference between what is desired, the target, and what can be achieved, the optima. A data center room with ten racks and one CRAC unit is considered with the total IT load in the room increasing from 10% to 100% of full load capacity uniformly in ten steps (over ten years). For each instance (year), the optimization framework informs distribution of IT load among the ten racks (for a given total IT load

for that year) and CRAC blower speed such that cooling energy consumption and deviation in rack inlet temperature from reference value is minimized. It should be noted that the CRAC supply temperature is not calculated by the optimization framework but by assuming a linear relationship between CRAC supply temperature and rack inlet temperature. To gauge energy savings for each year, the cooling energy consumption for optimized design is compared with cooling energy consumed for a baseline case wherein IT loads are randomly distributed and the CRAC flow rate is calculated using overall energy balance. It was found that the optimized configuration resulted in energy savings of 11–45% for each year.

1.6.2 Dynamic Optimization

One approach to energy management and dynamic optimization in data centers employs dynamic IT workload migration or placement to minimize energy consumption while maintaining the same level of performance [112-116]. This approach in most cases does not manipulate or account for facility-side parameters, so while such approaches have successfully minimized energy costs relating to IT workload assignment and processing, they do not include a significant portion of total energy consumption (cooling cost). As such, review of these studies is not included here; please see [117] for more details.

A thermally aware, power optimization framework is explored in [119] at both the server (involving a trade-off between fan power and circuit leakage power), as well as the data center (with a trade-off between server fan power and HVAC power consumption) levels. The cooling infrastructure and server power consumption are estimated using thermodynamics-based models and empirical curve-fitting, respectively. It is assumed that

all the servers are of the same design (IBM Power 750) and have the same utilization level. Energy savings are demonstrated by simulating a simple test case of implementing a binary control method that chooses between two thermal set-points according to the utilization level and cooling energy consumption in the data center. Tests for a single rack demonstrated a reduction in power consumption by as much as 12.4−17%. It should be noted that the rack inlet temperature is not explicitly modeled in this study, and compliance with temperature thresholds is only verified in case of change in server utilization levels or cooling set-points.

Model Predictive Control (MPC) based thermal management with cooling cost minimization for data center energy management have been explored by Zhou et al. [120, 121]. Temperature prediction at the rack inlet is obtained via a state-space linear model relating rack inlet temperature at the next instance to the current rack inlet temperature, CRAC supply temperature, and blower speed. Inherent assumptions in the formulation include uniform rack inlet temperatures for any given rack (single temperature point), and fixed airflow and recirculation patterns for the room configuration, which may not be true for different IT load distributions, and therefore server fan speed settings. The objective function for optimization is comprised of CRAC blower power consumption and chiller power consumption (assumed to be linear) and is minimized using a constrained minimization approach. In an effort to achieve control over local distribution of cooling resources, Adaptive Vent Tiles (AVT) were included in the optimization framework as an extension to previous studies. Depending on the scale of the implementation, AVTs are manipulated individually or as a group when populating a given cold aisle. The manipulated variables in the different studies include CRAC blower speed, CRAC supply

temperature and a combination of the three. The framework employing all three manipulated variables is implemented in an experimental area populated by 17 racks, and energy savings as great as 36% were demonstrated over a variety of test scenarios.

Chen, et al. [124] developed a Predictive Thermal and Energy Control (PTEC) system for data centers employing real-time temperature prediction algorithm [60]. Empirically developed correlations were used to estimate the power consumption of the air conditioning (AC) unit as a function of set-point temperature, blower speed and return air temperature, as well as the power consumption of the server fans as a function of server fan speed. The framework uses a Constrained Simulated Annealing (CSA) algorithm to search for cooling set-points (from among six set-point temperature and blower speed combinations) that minimize cooling power consumption while maintaining thermal safety requirements. CSA is a non-gradient based, sequential optimization search technique for solving constrained global optimization problems. The performance of the developed framework is evaluated on a test bed consisting of 15 1U servers in a single rack experimentally and on a data center model consisting of single cold aisle (total 229 servers) using CFD simulations. Estimated power consumption when PTEC is employed is compared with power consumed in a baseline case (static set-points resulting in overcooling) resulting in approximately 34% reduction in cooling power consumption. An important assumption for the temperature prediction model is that all the servers in the data center room are identical and have uniform CPU utilization.

1.7 Scope of Dissertation

The aim of this doctoral research is the development of a framework for allocation of cooling resources in a data center to ensure reliable operation while minimizing cooling energy consumption. This framework consists of several components and is illustrated in Fig. 9.

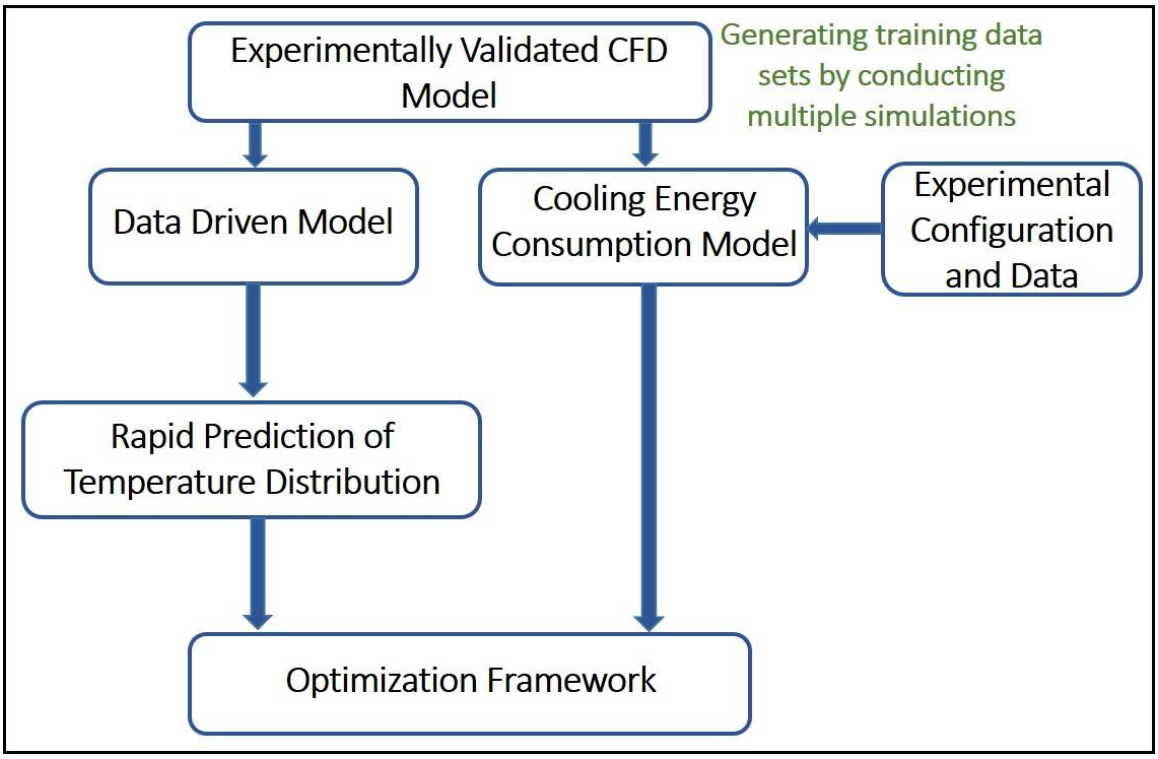


Figure 9 - Developed Framework

The first section of this dissertation (Chapters 2, −4) focuses on modeling of data centers. Specifically, the objective was development of models for rapid prediction of temperature distribution and flow rates in data centers which could potentially be employed in dynamic control framework. Data driven modeling frameworks were explored for this purpose. The training data for the data driven models were obtained by conducting simulations using numerical model for the data center room under consideration. Chapter

2 presents details of room-level CFD/HT model developed using software Future Facilities 6SigmaRoom. The CFD/HT model developed is validated using experimental measurements. Chapters 3 and 4 present details of rapid models developed using four different data driven modeling frameworks: ANN, Support Vector Regression (SVR), Gaussian Process Regression (GPR) and POD. To the best of our knowledge, this is the first application of SVR and GPR techniques to thermal modeling and temperature prediction in data centers. Chapter 4 presents a comparison of performance of the four models, to provide guidance on selecting an appropriate data driven modeling approach.

Development of thermodynamics and heat transfer based cooling energy consumption model for data centers (involving modeling of CRAC and chiller system) is presented in Chapter 5. The form of the model developed is such that it can directly be integrated with the overall optimization framework. Chapter 6 presents development of a GA based optimization framework for cooling energy minimization while ensuring that rack inlet temperatures are maintained within prescribed thresholds. Two case studies implementing this framework were conducted and results are presented. Finally, future work, recommendations and conclusions are outlined in Chapter 7.

CHAPTER 2. ROOM LEVEL COMPUTATIONAL FLUID DYNAMICS MODELING OF DATA CENTERS

This chapter presents an experimentally validated room-level CFD/HT model for raised-floor data center configurations based on the commercial software package Future Facilities 6SigmaDCX [72]. As indicated in Chapter 1, CFD/HT modeling is extensively used to model thermal transport and air flow in data centers. Length scales in a data center can range from tens of meter for the entire room to tens of nanometer for individual transistors in processor chips; It is therefore difficult and prohibitively expensive in terms of computational cost to construct room-level models that include details of all components in a typical data center. Several CFD models have been developed for smaller test cells (*e.g.* the plenum, individual racks.) or individual components (*e.g.* individual servers, cooling and power components) to obtain detailed results at appropriate length scales (see Table 3). The results from these models are then usually used in turn to construct "reduced-order" models of the individual components that can be incorporated into room-level models to achieve good accuracy at reasonable computational cost. In [73], a method to determine a simplified server model (SSM) from detailed server modeling is illustrated. This simplified model, which consists of imposing appropriate boundary conditions at the server inlet and exit to accurately determine heat and flow parameters, was then utilized in a computationally efficient room-level model. Zhang et al. [74] studied the level of detail needed in rack modeling by comparing results from three models with increasing levels of

Reference: J. Athavale, Y. Joshi, and M. Yoda, "Experimentally Validated Computational Fluid Dynamics Model for Data Center with Active Tiles," *ASME Journal of Electronic Packaging,* vol. 140, pp. 010902-010902-10, (2018).

detail. They concluded that there is no significant difference between the results for the three models, and hence recommended using a computationally efficient "black box" approach for room-level modeling.

Additionally, as an extension to the baseline data center room configuration, which employs only passive tiles, a numerical model employing active tiles was also developed. Active tiles are perforated floor tiles with integrated fans (see Fig. 13(a)), which increase the local volume flowrate by redistributing the cold air supplied by the CRAC unit to the under-floor plenum. In a previous study, [1], steady-state and transient experiments were conducted using active tiles to characterize their effect on temperature and flow distribution in a data center room. The authors also explored how active tiles affected local flow conditions and could be used as actuators to dynamically modulate the cold air distribution from the plenum. They reported that active tiles, as the actuators closest to the racks, can significantly and quickly impact the local distribution of cooling resources. They could therefore be used in an appropriate control framework to rapidly mitigate hot spots, and maintain local conditions in an energy-efficient manner.

2.1 Data Center Lab Description

The Data Center Laboratory at the Georgia Institute of Technology is divided into two 56 m^2 "halves" as shown in Fig. 10. This study only considers the highlighted experimental zone on the left. This zone houses three CRAC units sited around the periphery, one PDU, and 12 racks, which are arranged on both sides of a single cold aisle. Of the 12 cabinets, nine are fully populated with servers of varying configurations and power densities as in a typical data center. Two have servers in approximately half the

slots, while blanking panels cover the remaining height of the rack at the inlet, and one is a server simulator containing heater banks. Details of all the IT and cooling equipment in the data center room can be found in Table A.1 in Appendix A. The plenum is 0.91 m deep, and facilitates uniform flow through the perforated tiles.

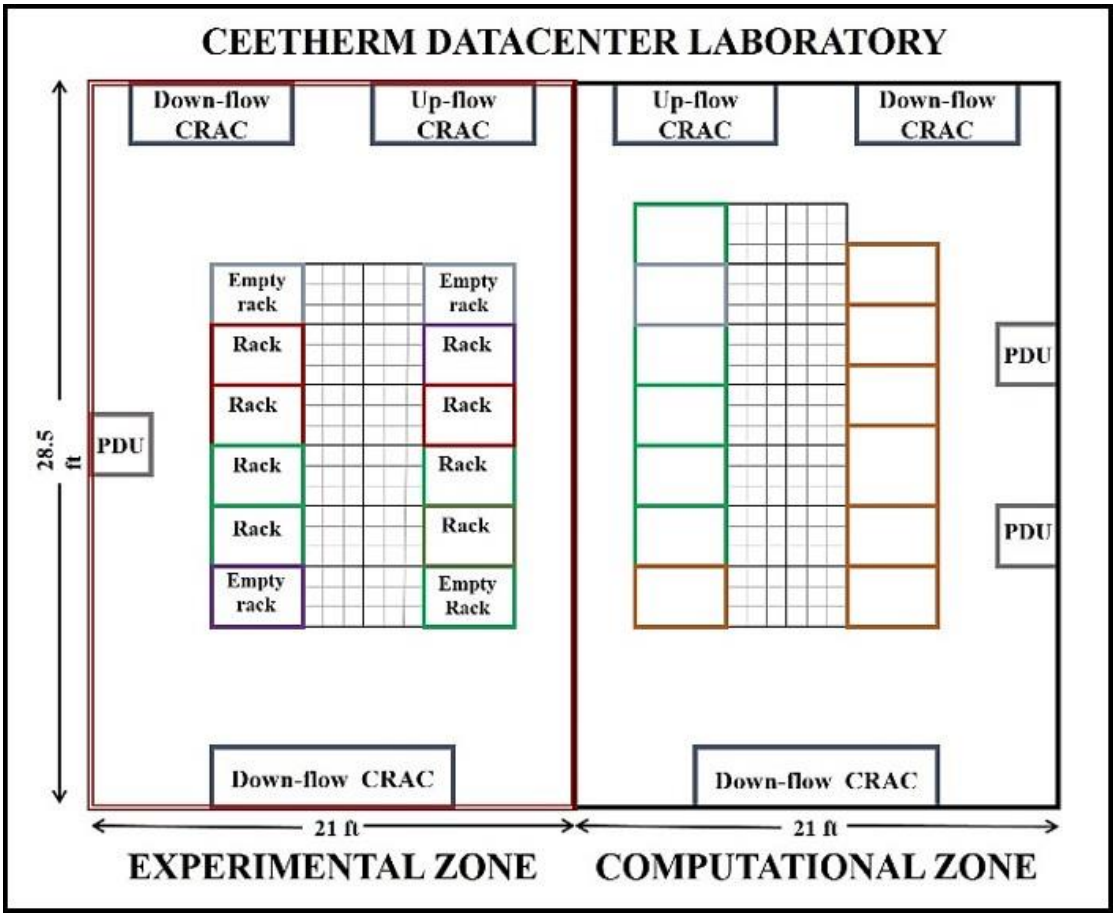


Figure 10-Data Center Experimental Facility

Figure 11 shows the naming scheme used to identify specific racks and tiles in this study, the pressure sensors, and the status of CRAC units. The plenum pressure sensor was located in the mid-plane of the cold aisle, and 0.4 m below the tiles, whereas the reference pressure is measured at the top of rack. For the experiments and simulations, the cold air was provided in an under-floor supply, ceiling return configuration.

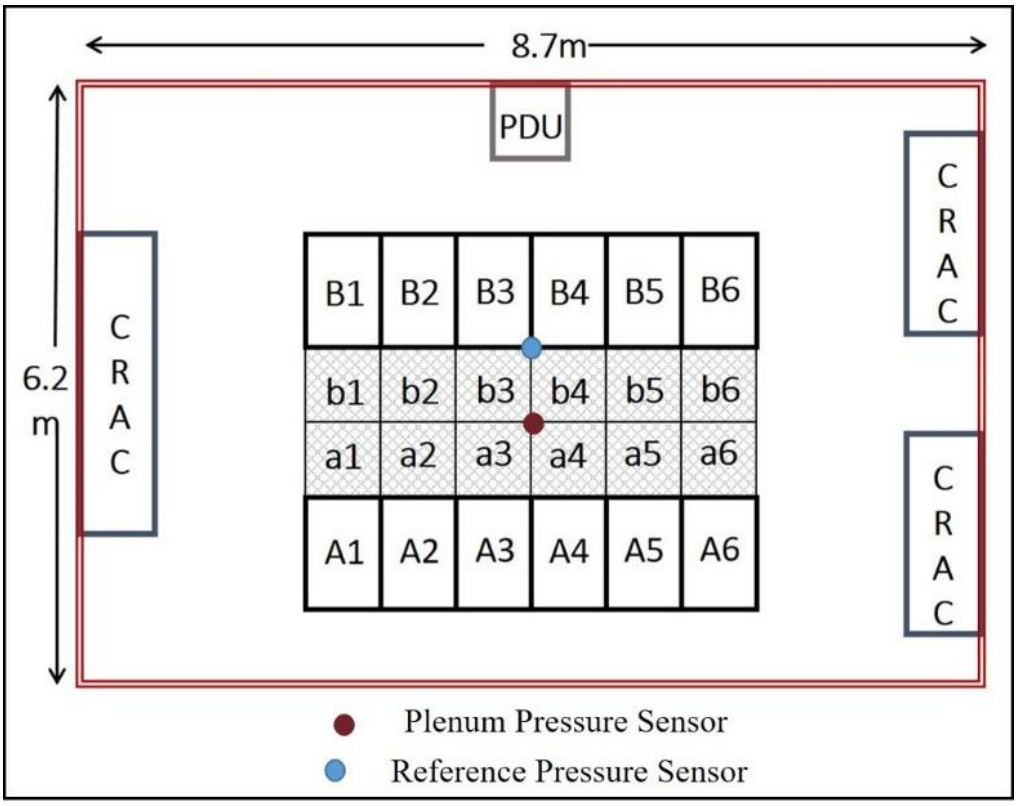


Figure 11 - Experimental Set-Up

2.2 Numerical Modeling

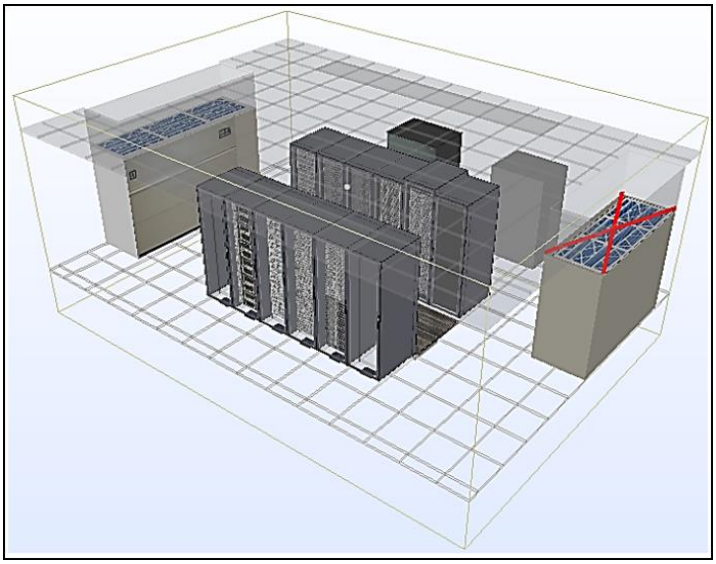
Computational fluid dynamics (CFD) analysis was carried out using the commercial package 6SigmaRoom [72], which is specifically designed for room-level data center simulations. The details for the numerical solver being employed can be found in [75]. The baseline model considers a configuration with all passive tiles. Experimental data for this configuration were used to calibrate plenum leakage and resistances for the model. Subsequently, a CFD model incorporating active tiles was developed for two configurations: (a) a single active tile and nine passive tiles in the cold aisle; and (b) an aisle populated with ten (*i.e.*, all) active tiles. Details for modeling different components in the three configurations are described below.

2.2.1 Baseline Model

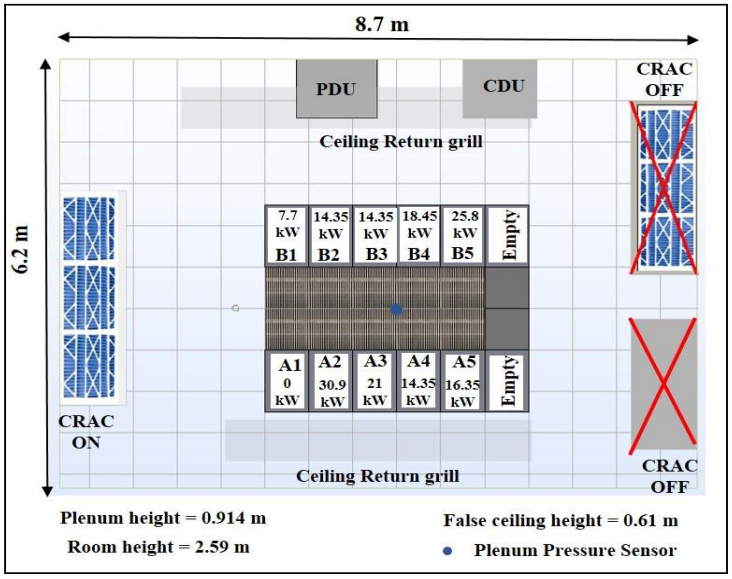
Figure 12 shows isometric (a) and plan (b) views for the baseline model for the data center room. In the previous study, it was experimentally determined that the active tiles under consideration have an effective porosity of 27% [53]. Hence, all the passive tiles are modeled to have a porosity of 27%. For all racks, except for the server simulator, characteristics of particular server designs have been included in the model, along with the control logic and fan curves for server fans to accurately model flow through racks, in response to rack air inlet temperature and pressure, respectively. For the sever simulator, specific fixed flow rate can be set, based on experimental conditions for different studies. Appropriate resistances in the plenum were estimated by comparing results from baseline model with experimental data, and these have been included in the model in the form of porous obstructions.

All three CRAC units in the room were modeled. However as only one CRAC unit was operational for all simulations, the return and supply openings for the other two units were blocked using solid obstructions to prevent air being passively introduced into the room or plenum space through these openings, mimicking the experimental set-up. Blower curves were used to model the relation between pressure and flow rate for the CRAC blowers. The operational CRAC unit is a Liebert FH740 [76], which has three blowers and a maximum airflow rate of 19,917 CFM when the blowers are running at 100% blower speed (1520 RPM) and back pressure of 0 Pa. A pressure sensor point for relative plenum pressure measurement was introduced in the numerical model at the same location as the experimental set-up for comparison with experimental results. The PDU is cooled using

air from the plenum and is modeled to have the same inflow and outflow area for cold air as in the actual facility.



(a)



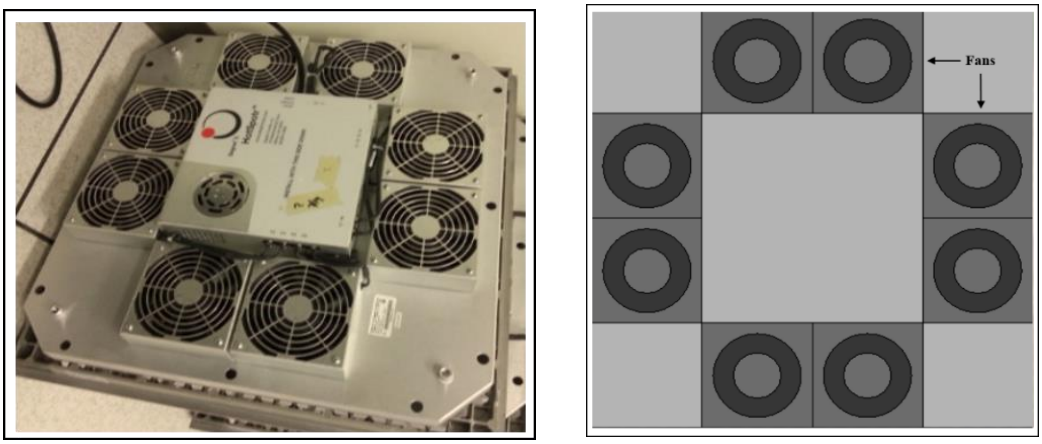
(b)

Figure 12 - (a) Baseline Model – Isometric View (b) Baseline Model – Plan View

2.2.2 Model with Single Active Tile

This configuration models the flow and temperature distribution in an aisle with single active tile and other passive tiles, and hence is a typical configuration where active

tiles are used to mitigate isolated hot spots. Figure 13(a) shows an active tile with eight integrated fans, each with a maximum flow rate of 171 CFM, at rated rpm of 3,600. In the numerical model, the active tile consisted of eight fan elements below a perforated grill having a porosity of 56%, as specified by the manufacturer [77]. Solid obstructions were used to model the central region and the four corners, as shown in Fig. 13(b). The different elements were arranged to closely mimic the actual active tile shown in Fig. 13(b). Although the tile fans operated at a fixed speed for a given simulation, fan curves for different fan speeds were incorporated in the model to capture the effect of plenum pressure on flow rate through the tile. Fig. 14 shows the CFD model with a single active tile, at location A3 in front, and at the base, of the server simulator.



(a) **(b)**

Figure 13(a) - Active Tile (b) - Numerical Model Representation for Active Tile

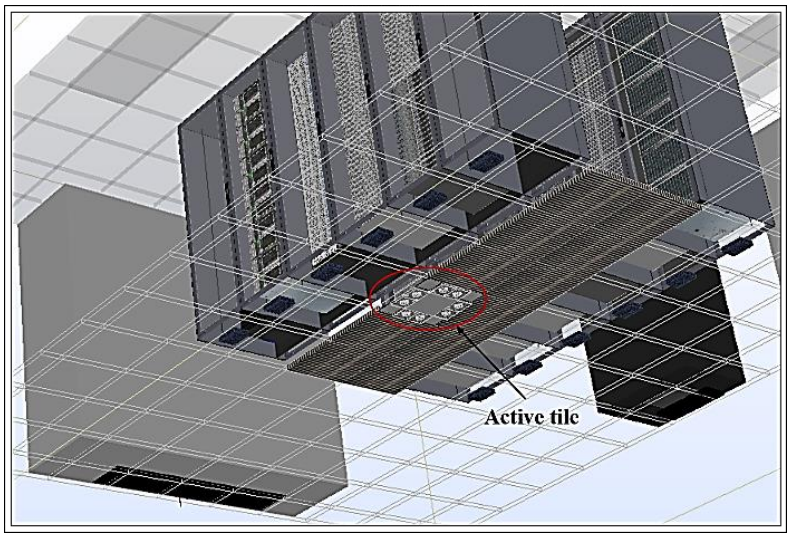


Figure 14 - Model with Single Active Tile

2.2.3 Model with Aisle of Active Tiles

Figure 15 shows the numerical model of the configuration with an aisle of active tiles, modeled as described before.

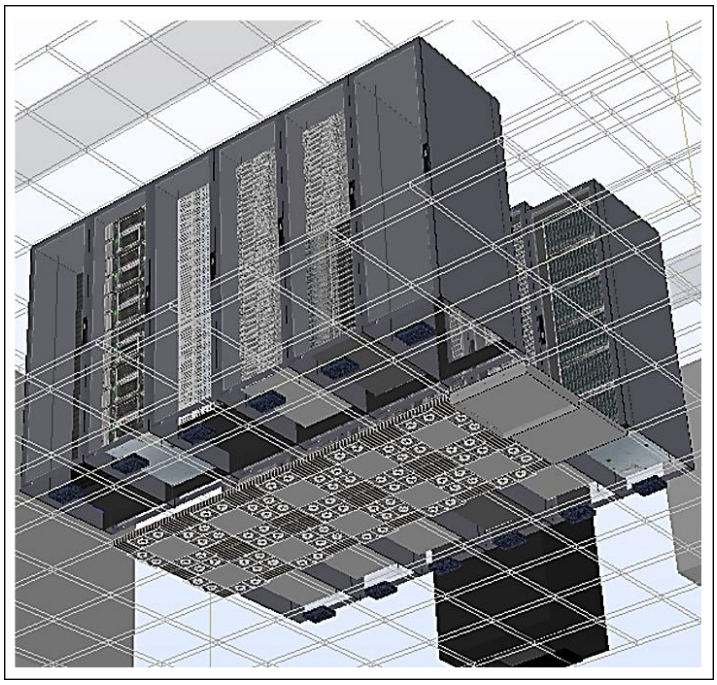


Figure 15 - Model with Aisle of Active Tiles

2.2.4 Computational Set-Up

The domain is discretized using a globally uniform structured grid of hexagonal elements. Inflation layers of five grid cells were created on top and bottom of solid objects to more accurately simulate the heat transfer between the solid and air. A grid independence study was conducted for each configuration of the model. The total flow rate through all the tiles and average rack inlet temperature for select racks were compared over 12 successively finer grids, with the total number of elements varied over two orders of magnitude, from 0.3×10^6 to 30×10^6. Table 6 gives the number of grid elements used to model each configuration, chosen so that the maximum discrepancy in total flow rate through the tiles and average rack inlet temperature for rack A3, between the grid used and the grid with 30×10^6 elements was less than 0.2% and 1.6%, respectively.

Table 6 - Grid Selected for Different Models

Numerical Model Configuration	Number of grid elements
Baseline Model (All Passive Tiles)	2.26×10^6
Model with Single Active Tile	2.75×10^6
Model with Aisle of Active Tiles	4.14×10^6

Data center rooms are characterized for the most part by turbulent flow conditions, and the standard k-ε model was used to model the turbulence by solving the additional equations for turbulent kinetic energy and dissipation rate. 6SigmaRoom uses the finite-volume method to discretize the Reynolds-averaged Navier-Stokes equations, along with the energy equation. The velocity and temperature fields are solved in the flow domain and on the fluid-solid interfaces; the solid structures like servers and CRAC units are not simulated. The solution is assumed to be converged when the residuals of the variables between consecutive iterations is less than 10 ppm.

2.3 Results and Discussion

2.3.1 Experimental Validation

2.3.1.1 Baseline Model

The baseline model consists of a complete aisle populated only with passive tiles. Figure 16 compares the experimentally measured and numerically predicted tile flow rates as a function of tile location for a CRAC supply temperature and blower speed of 15.5 °C and 80%, respectively, and uniform rack heat load of 10 kW for each rack.

The tile flow rate was measured using a grid of thermal anemometers, as shown in Fig. A.1 in Appendix A [78], and detailed in [53]. The air velocity measurements have a manufacturer-specified uncertainty of ±5%, which was confirmed using a handheld anemometer. The maximum relative discrepancy between the CFD and experimental results is 5.2%, and the tile flow rate predicted by the model is generally less than the experimentally measured value (except for tile A3). The simulation results over a range of blower speeds show that the maximum discrepancy usually occurs for either tiles A2 or A3. The data center room has a sparse honeycomb flow straightener located in the plenum at a depth of 0.3 m below the floor at this location (below A2 and A3), which was not included in the numerical model. We suspect that this explains the larger discrepancies for tiles A2 and A3, and the larger tile flow rate predicted for tile A3.

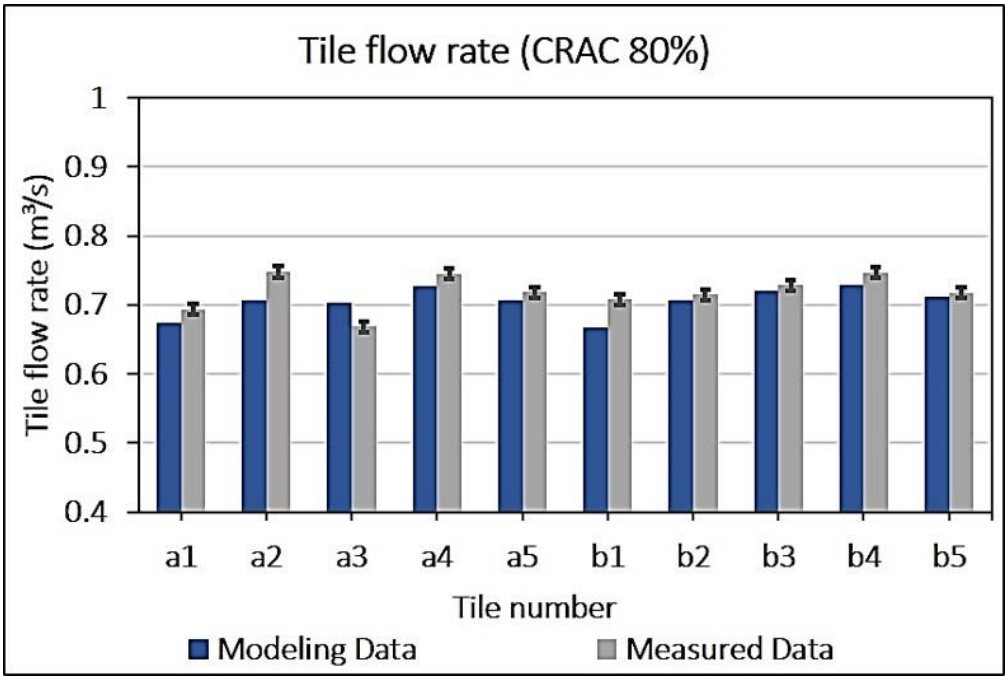
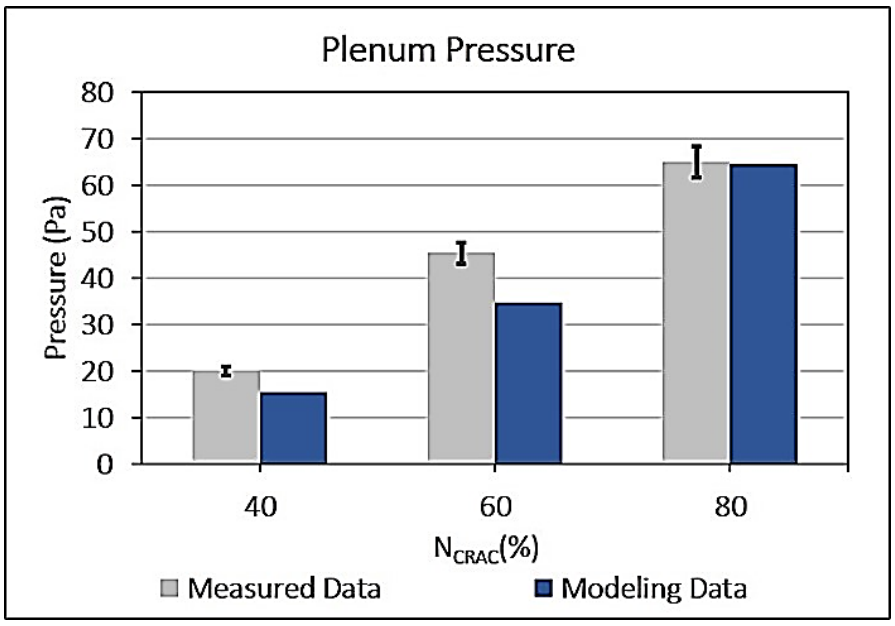
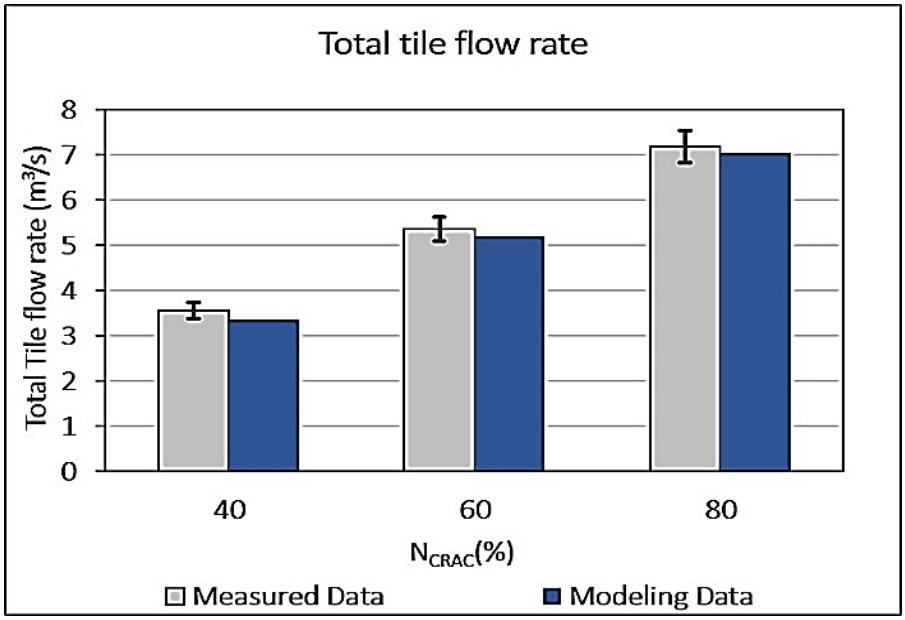


Figure 16 - Comparison of Measured and Predicted Tile Flow Rate for Baseline Case



(a)



(b)

Figure 17(a) - Comparison of Measured and Predicted Relative Plenum Pressure for Different CRAC Blower Speeds (b) - Comparison of Measured and Predicted Total Tile Flow Rate for Different CRAC

Figure 17 compares the total flow rate through all the tiles (a) and relative plenum pressure (b) as a function of CRAC blower speed. The pressure was measured by an Alnor micromanometer AXD610 [79] with a manufacturer-specified uncertainty of ±1%, as

detailed in [53]. The numerical predictions and experimental measurements of relative plenum pressure, and consequently total flow rate through the perforated tiles are in very good agreement. The simulations show that the total flow rate through the cold aisle is a strong function of relative plenum pressure, and accurate prediction of plenum pressure will therefore give an accurate prediction of total tile flow rate, despite discrepancies in individual tile flow rates. Although the plenum pressure is based on a value at a single location in the experiment and the simulations, the pressure is quite uniform over the plenum, as shown in numerical predictions of the pressure field over a horizontal plane 0.4 m below the floor grill (= depth of the pressure sensor) for a CRAC blower speed of 60% shown in Fig. 18.

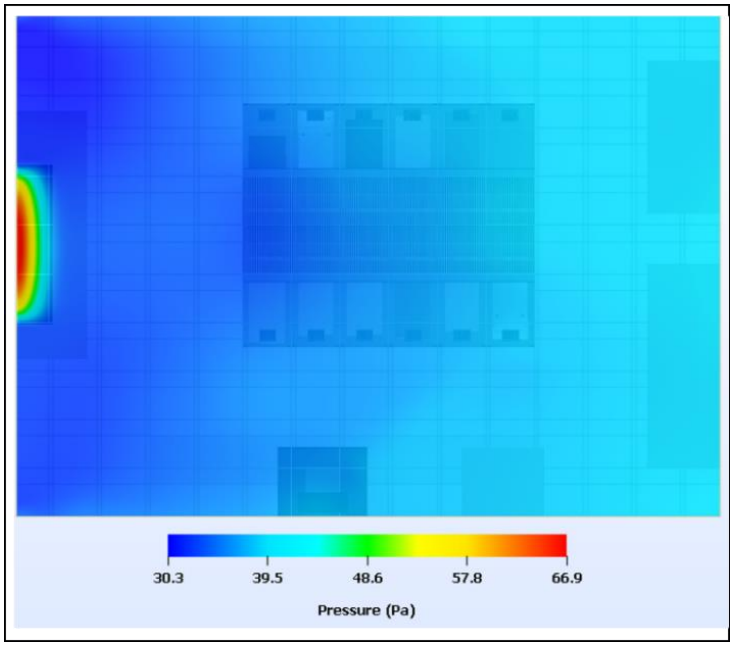


Figure 18 - Plenum Pressure contour for CRAC Blower Speed of 60%

2.3.1.2 Model with Single Active Tile

Figure 19 compares the measured and modeled tile flow rates at tile A3, as a function of active tile fan speed, for a CRAC blower speed of 60%, supply temperature of

15.5 ℃ and rack heat load of 10KW for each rack. In this figure, an active tile fan speed of 0% corresponds to the baseline case of passive tiles. The maximum relative discrepancy between the measured and modeled values is 4.6%.

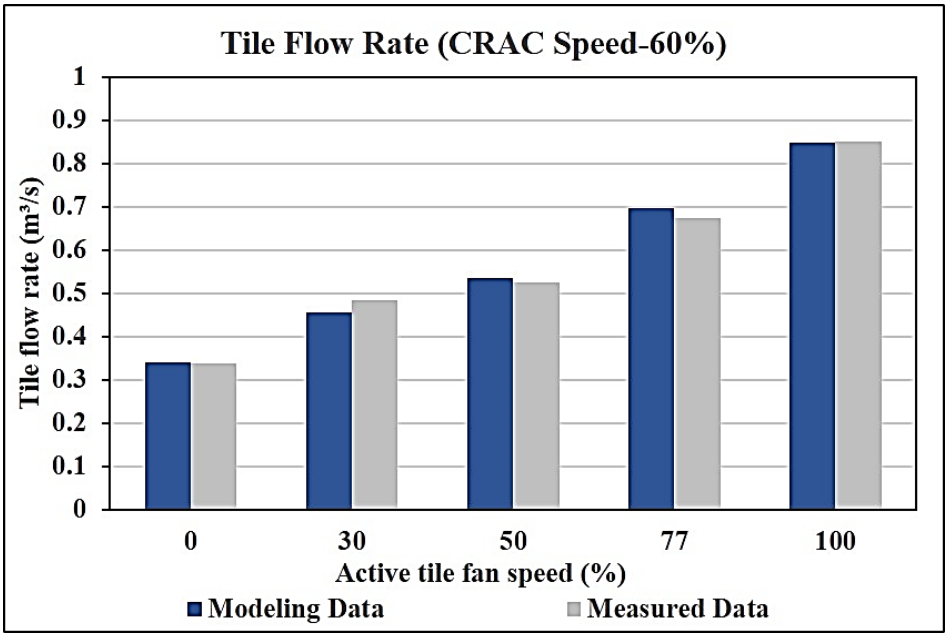
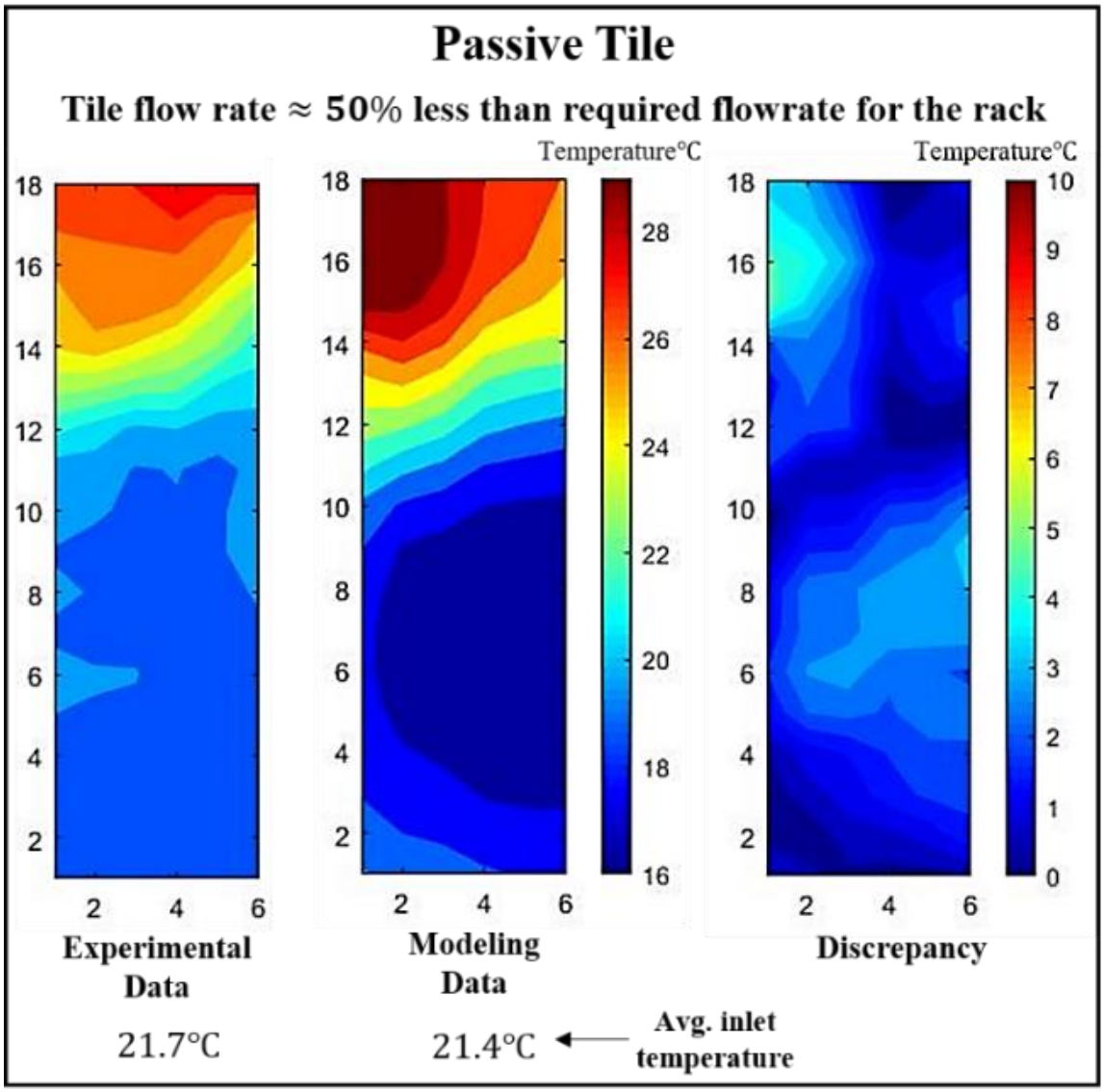


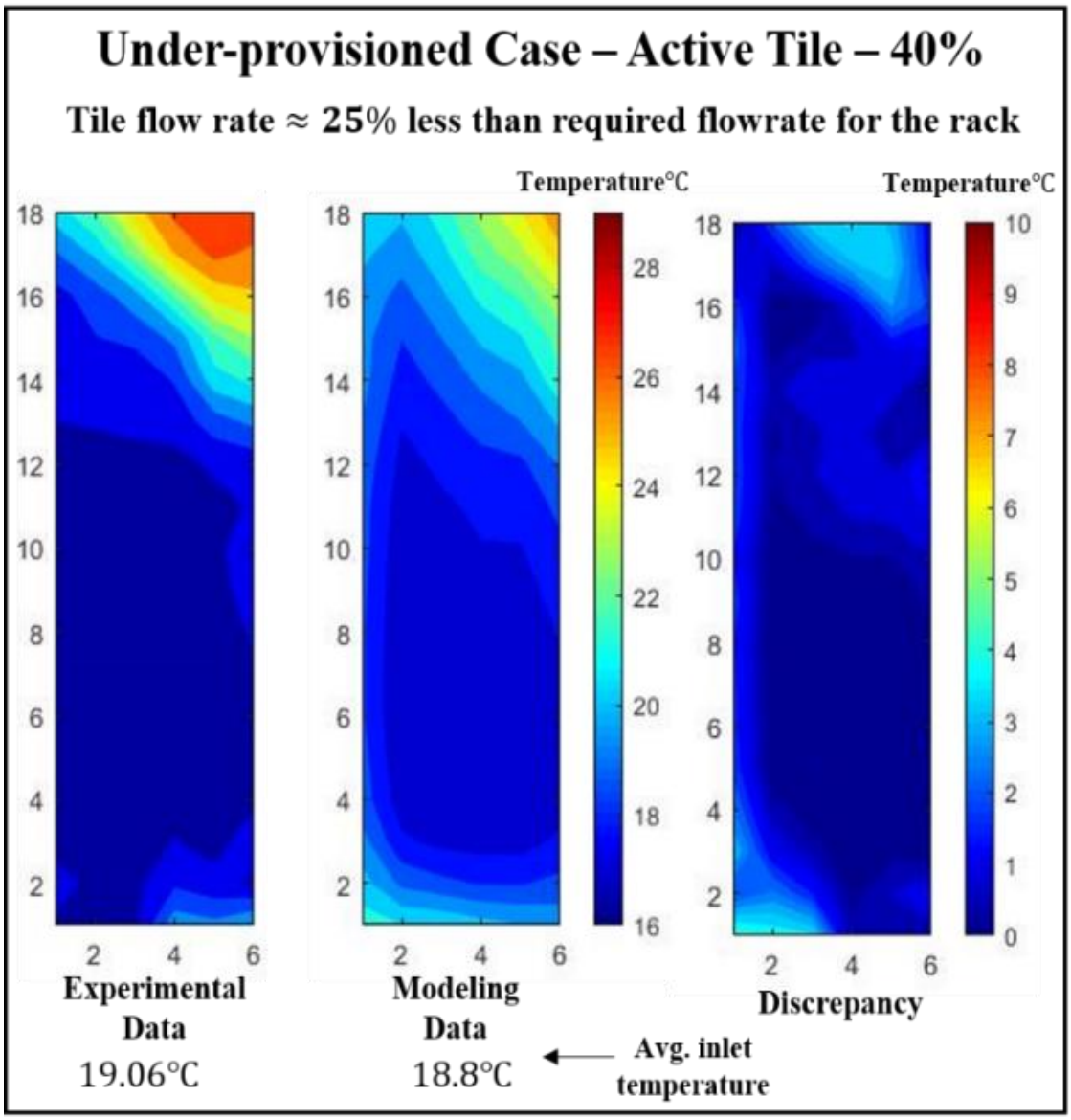
Figure 19 - Comparison of Measured and Predicted Tile Flow Rate for Different Active Tile Fan Speeds

Simulations were also performed to compare the measured and predicted rack inlet temperature profiles for different active tile fan speeds at a CRAC blower speed of 60%. The temperature profiles in the experiments were obtained using a grid of 108 T-type thermocouples (TC) (see Fig. A.3 in Appendix A) with a measurement uncertainty of ±0.5 ℃, and bilinear interpolation between the measurement points [53]. The results at the 108 TC locations were directly compared with numerical results at the same locations. Rack A3 (server simulator) is set to have a heat load of 10 kW, while drawing air at a constant flow rate of 0.63 m^3/s. Two different scenarios were considered, corresponding to whether the server simulator is under- or correctly provisioned, using a single active tile,

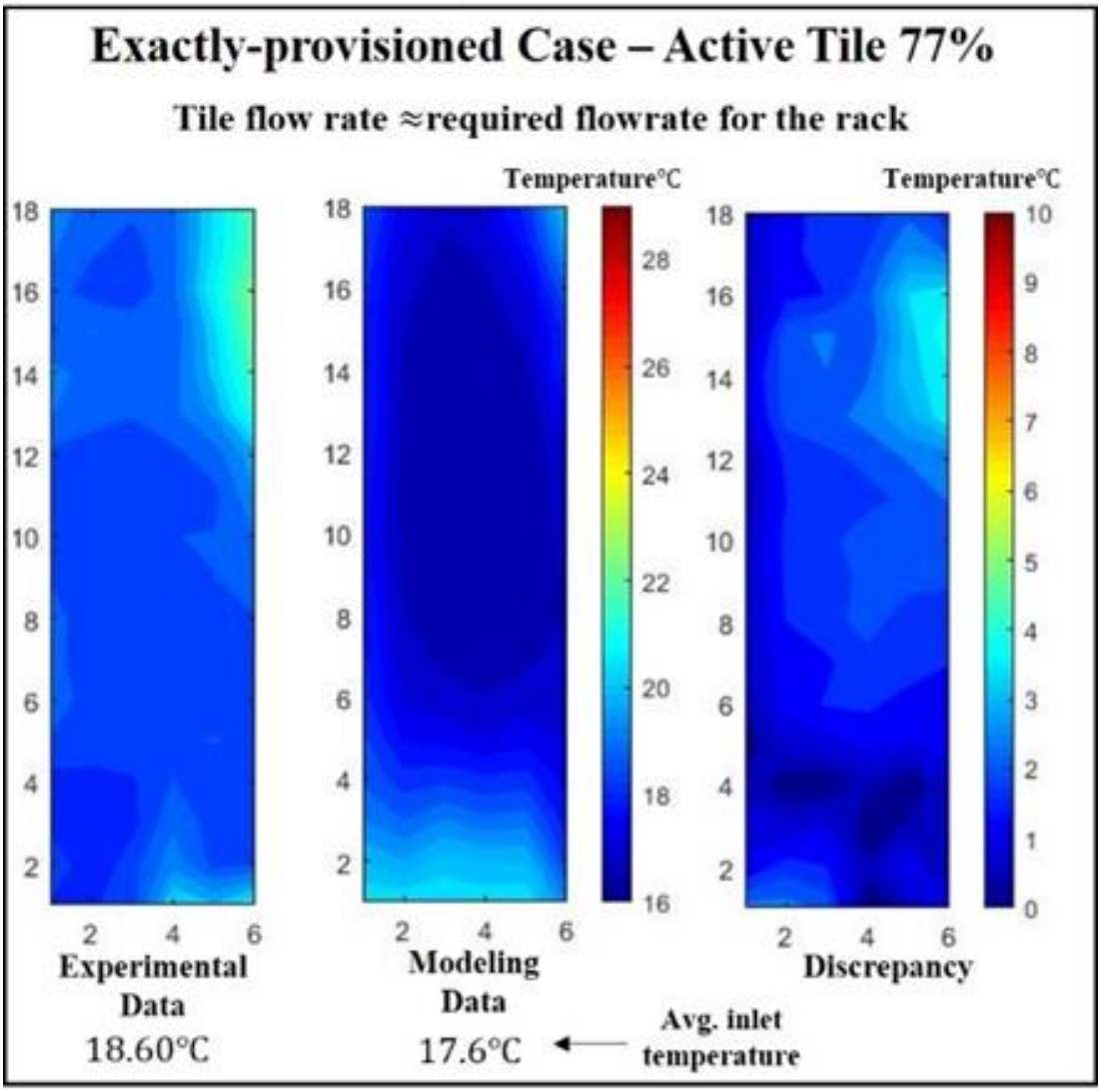
and the results were compared to the baseline case with passive tiles at the same CRAC blower setting.



(a)


Under-provisioned Case – Active Tile – 40%
Tile flow rate ≈ 25% less than required flowrate for the rack
Temperature°C
Temperature°C
Experimental Data
19.06°C
Modeling Data
18.8°C
Avg. inlet temperature
Discrepancy

(b)



(c)

Figure 20 – (a) - Comparison of Measured and Predicted Rack Inlet Temperature Contour for Baseline Case (b) - Comparison of Measured and Predicted Rack Inlet Temperature Contour for Under-Provisioned Case (c) - Comparison of Measured and Predicted Rack Inlet Temperature Contour for Exactly-Provisioned Case

Figure 20 shows the experimentally measured temperature field, the numerical predictions, and the difference between the two for the three scenarios considered. The x and y axes mark the locations of the temperature sensors used for measurement. For the baseline case with passive tiles, (a), and the under-provisioned scenario with active tile fan

speed of 40%, (b), the numerical model successfully captures most of the qualitative temperature trends, with a modest discrepancy in actual temperatures. For the case with passive tiles, the simulations give a conservative prediction of rack inlet temperatures. The average inlet temperature for all cases, is also given in the Fig. 20.

2.3.1.3 Model with Aisle of Active Tiles

This section compares experimental and numerical results for entire aisle of active tiles where the fan speed is the same for all the fans for all the tiles. Figure 21 shows the individual tile flow rate for a CRAC blower speed of 80% and active tile fan speed of 77%. For most tile locations, the numerically predicted flow rate is greater than the experimentally measured value.

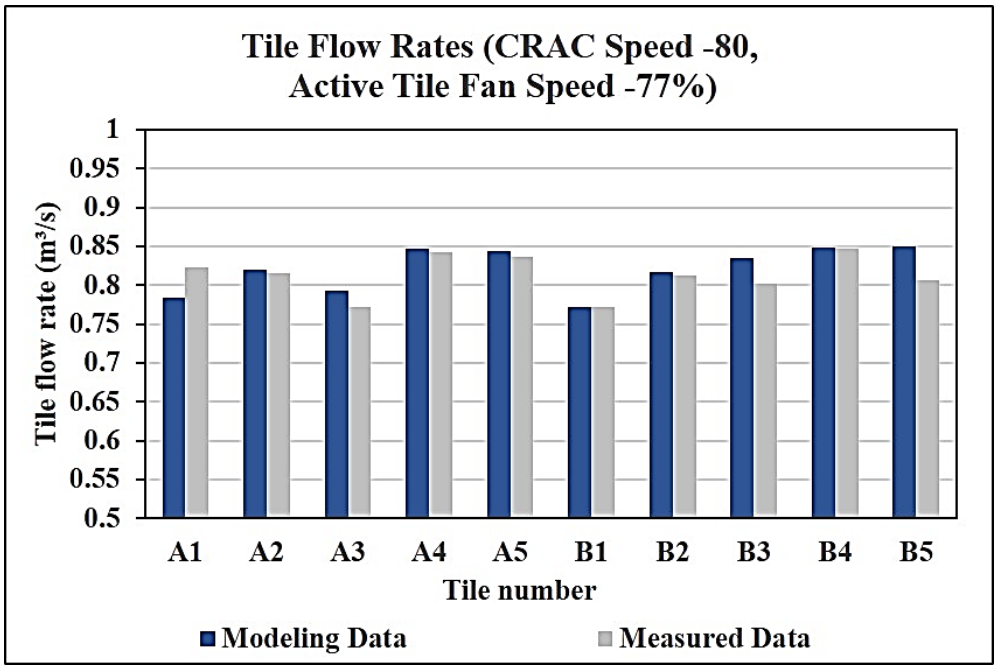


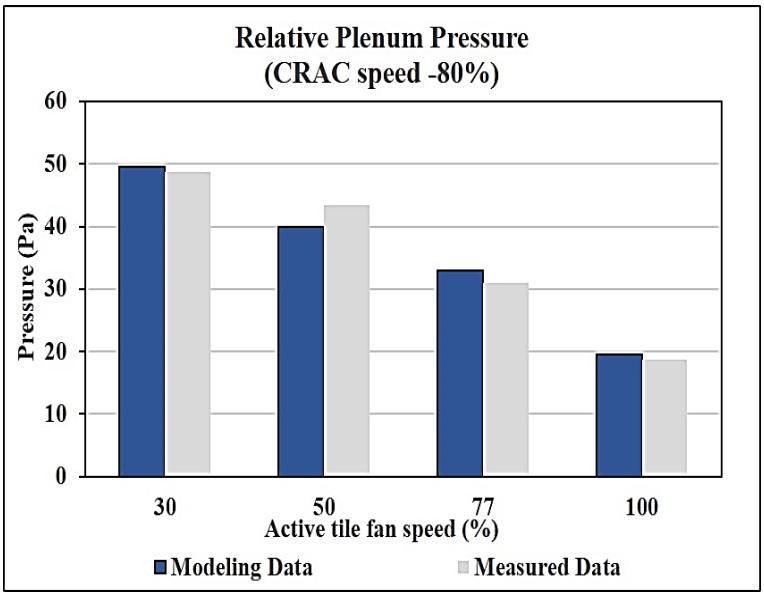
Figure 21 - Comparison of Measured and Predicted Tile Flow Rate for an Aisle of Active Tiles

Figure 22 compares the relative plenum pressure (a) and total flow rate (b) as a function of tile fan speed at a CRAC blower speed of 80%. Because the active tile fans are

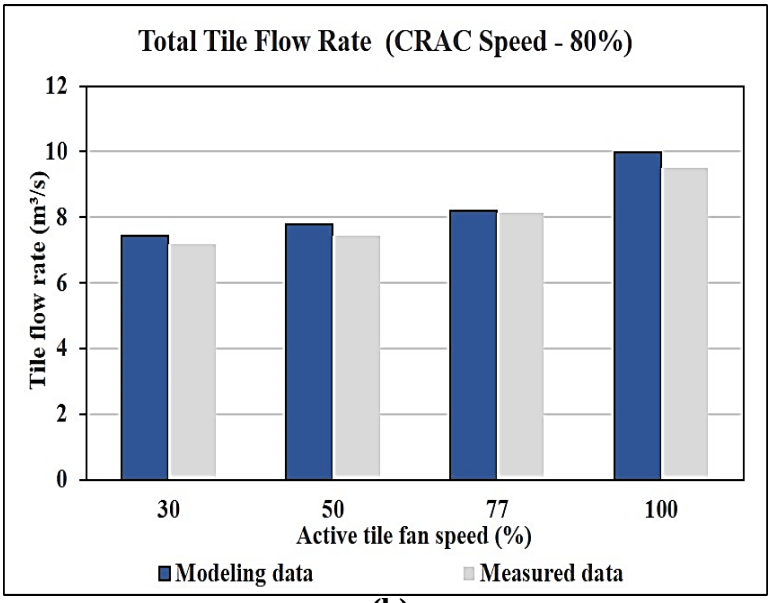
in series with the CRAC blower, there is a decrease in the pressure between the two, which leads to a decrease in the relative plenum pressure as the active tile fan speed increases—a trend captured by the numerical model in Fig. 22(a). The maximum relative discrepancy in total tile flow rate between the simulations and experiments is 4.4%, while that in relative plenum pressure is 3.4 Pa. Table 7 gives the average discrepancies in flow rate, temperature, and pressure measurement between the three numerical models and experiments.

Table 7 - Average Discrepancy for Numerical Models Developed

Average Discrepancy→ Model ↓	Total Flow Rate (Relative) (%)	Pressure Value (Pa)	Temperature Value (℃)
Baseline Model	3.4	5.2	1.6
Model with Single Active Tile	3.8	-	1.6
Model with Aisle of Active Tiles	3.3	1.8	-



(a)



(b)

Figure 22 - (a) Comparison of Measured and Predicted Relative Plenum Pressure for different Active Tile Fan Speeds (b) - Comparison of Measured and Predicted Total Tile Flow Rate for different Active Tile Fan Speeds

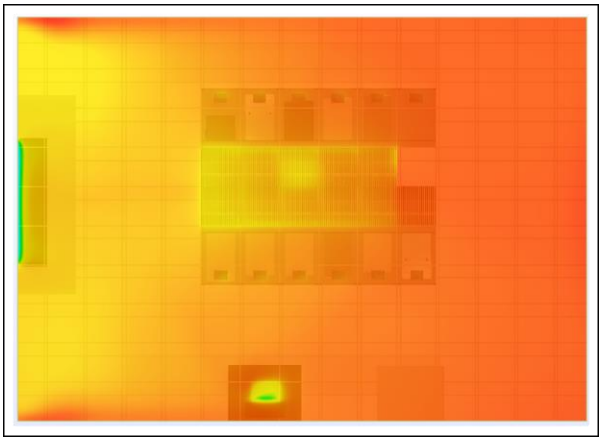
2.3.2 Numerical Model Results

The experimentally validated models are then used to study the temperature and flow patterns in data centers under different operating conditions. The models can provide

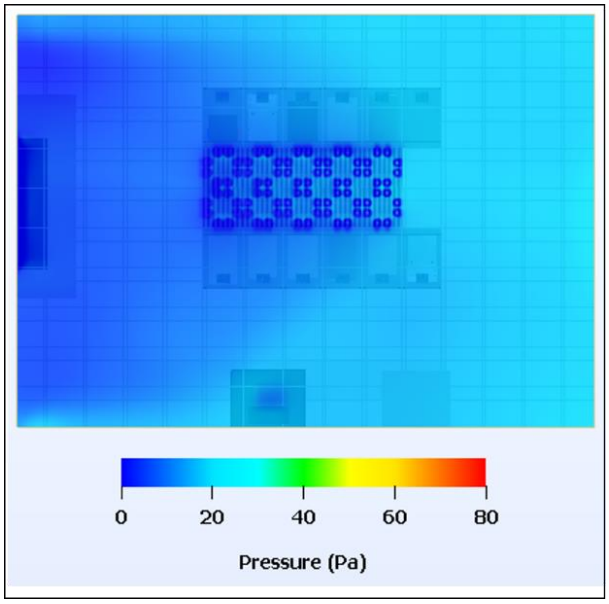
results with finer spatial resolution than experimental measurements, and can also be used to analyze cooling failure scenarios, which cannot be experimentally studied without risking damage to the IT and/or cooling infrastructure. The following sections illustrate some case studies that were simulated with the models.

2.3.2.1 Pressure and Velocity Contours

Pressure and velocity contours for planes very close (0.07 m below) to the tiles in the plenum space are presented in Figs. 23 and 24 for the baseline model and the model with an aisle of active tiles, respectively for a CRAC blower speed of 80%, and active tile fan speed of 100%. As mentioned previously, for the same CRAC blower speed, replacing passive tiles by active tiles decreases the pressure in the plenum (Fig. 23). It is noted that the low-pressure areas formed below each of the active tile fans in Fig. 23(b) correspond to a high-velocity region in Fig 24(b). Simulations show that the vertical component of air velocity through active tile fans is approximately quadruple that through the passive tiles.

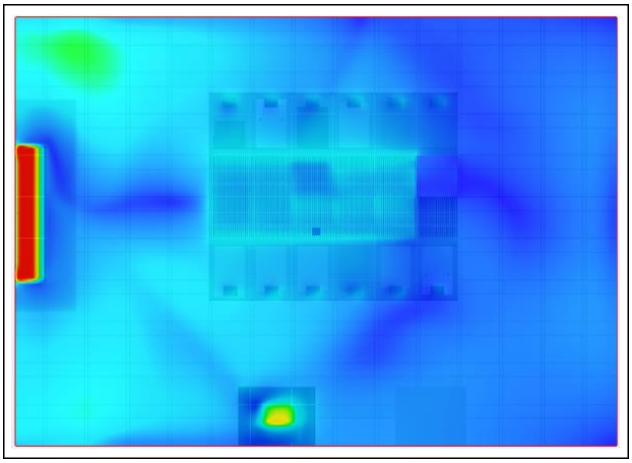


(a)

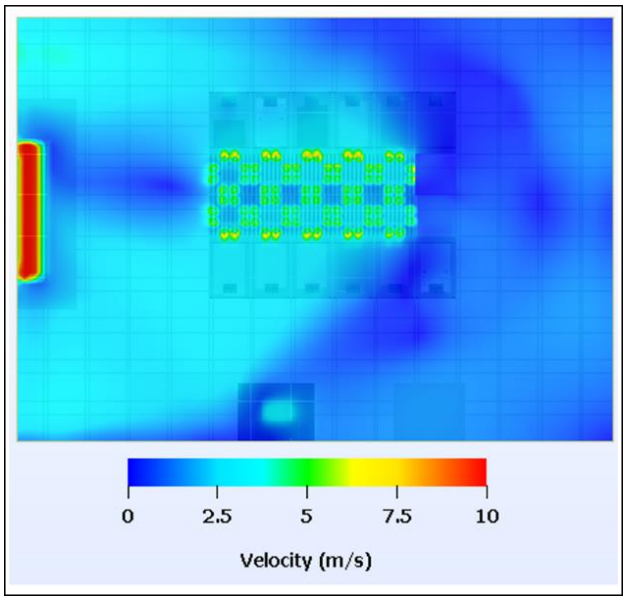


(b)

Figure 23(a) - Pressure Contour for Baseline Configuration (b) - Pressure Contour for Configuration of Aisle of Active Tiles



(a)



(b)

Figure 24(a) - Velocity Contour for Baseline Configuration (b) - Velocity Contour for Configuration of Aisle of Active Tiles

2.3.2.2 Effect of Active Tiles on Plenum Pressure

An ideal-case scenario in data center operation would be to have a small, but positive, relative plenum pressure, since this would minimize air leakage from the plenum to the data center room. For a given CRAC blower speed, relative plenum pressure decreases as the active tile fan speed increases for an aisle populated by active tiles,

resulting in less (undesirable) air leakage from the plenum to the data center room, and most of the air exiting through the perforated tile. Figure 25, which plots the fraction (in percent) of the total air supplied by the CRAC that translates to air leakage as a function of active tile fan speeds for different CRAC blower speeds, illustrates this phenomenon. At a low CRAC blower speed of 40%, a high active tile fan speed (77% or 100%) leads to a negative relative pressure in the plenum, resulting in room air being drawn into the plenum, a phenomenon also observed in the experimental study [53]. A validated CFD model can predict such undesirable situations, and determine where to locate active tiles to avoid this.

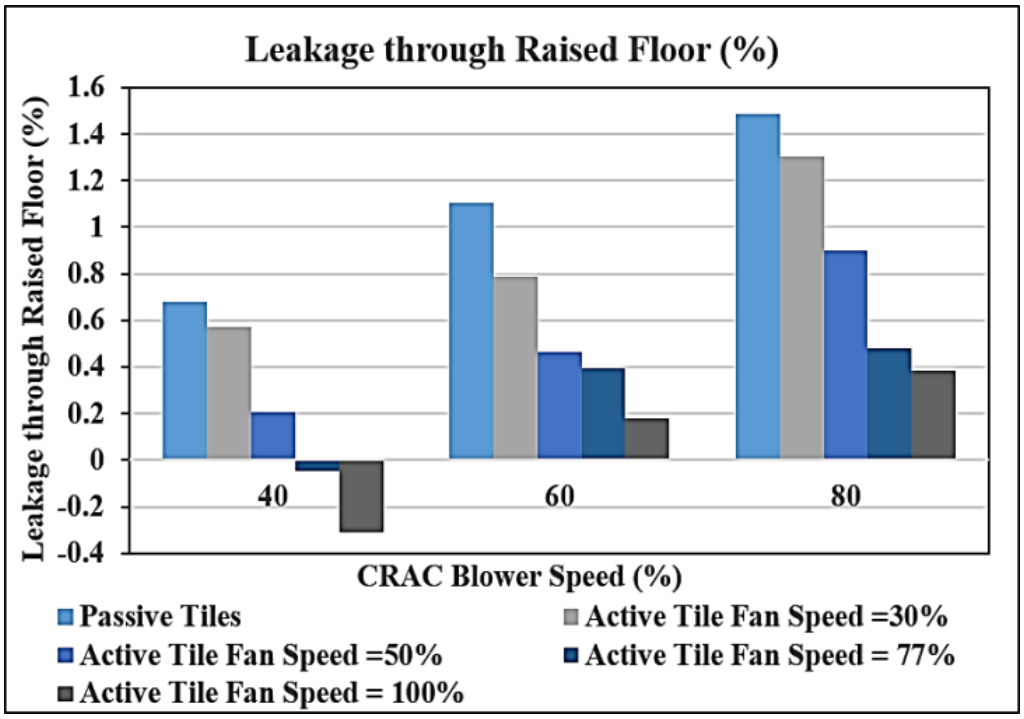


Figure 25 - Percentage Leakage through Raised Floor as a Function of CRAC Blower Speed and Active Tile Fan Speed

2.3.2.3 Failure Scenario

Data centers are classified as mission-critical facilities and as such, it is important to ensure that their operating conditions are in accordance with ASHRAE

recommendations at all times. A cooling failure in a data center can occur due to a power, or mechanical component, failure. During a cooling disruption, maximizing the time period before the IT equipment exceeds ASHRAE recommended temperature threshold and is at risk of failure or damage due to overheating, the ride-through period, is desirable because it gives data center operators more time to implement corrective measures.

In this study, two distinct failure scenarios were considered for both the baseline configuration, and the configuration with an aisle of active tiles. Scenario (A) represents the case where the cooling coil fails (chilled water pump failure), while (B) represents the case when CRAC blower fails. The objective of this study was to determine if employing active tiles would provide any advantage over passive tiles in terms of increasing the ride-through period for a cooling disruption. The simulations were first used to predict steady-state results in the absence of any failure, and then the failure of either the chilled water pump (A) or the CRAC blower (B) was simulated at time t = 0 s. For the steady-state simulations, the CRAC blower speed, CRAC supply temperature and active tile fan speed were 80%, 15 °C and 100%, respectively. Table 8 gives the ride-through times based on the simulations for any of the IT equipment to reach a maximum rack inlet temperature exceeding 32 °C, which violates the ASHRAE prescribed recommended temperature threshold for the different scenarios.

Table 8 shows that active tiles increase the ride-through time for both failure scenarios. This may be due to an increase in thermal mass available in the room due to the tiles, and/or circulation of air by the active tile fans. For passive tiles, the critical inlet temperature occurs earlier for scenario B when the CRAC blower fails than for A, when

cooling coil (chilled water pump) fails. This is expected because when the cooling coil fails, the CRAC blowers will still circulate the reservoir of cool air available in the plenum, keeping the IT equipment cooler for a longer period of time. However, if the fans fail, there is almost no active motion of air to carry the heat from the IT equipment and dissipate it to the cooling coil, even though the air is still being cooled.

Table 8 - Ride Through Time for Failure Scenario (A) and (B)

	Time to reach critical temperature (s)	
Failure Scenario→ / Configuration ↓	(A) Cooling Failure	(B) CRAC Blower Failure
Baseline	118	109
Aisle of Active Tiles	145	158

Conversely, the configuration with active tiles has a longer ride-through time for scenario B (CRAC blower failure) than A (cooling coil failure). In B, the active tile fans are still running and there is some recirculation of air which can absorb heat from the IT equipment and exchange it with the chilled water heat exchanger. This results in a longer ride-through time than A, because there is no cooling available to dissipate heat, even if the CRAC blowers are still functioning. In essence, this shows that a combination of air motion, albeit reduced, and functional chilled water pump can increase ride-through time.

Table 9 - Maximum CRAC Inlet and Exit Temperature for Failure Scenario (A) and (B) at t = 300s

	Failure Scenario (A) Cooling Failure	
Configuration ↓	Maximum temperature at CRAC inlet (℃) at t=300s	Maximum temperature at CRAC outlet (℃) at t=300s
Baseline	32.7	32.6
Aisle of Active Tiles	31.3	31.2
	Failure Scenario (B) CRAC Blower Failure	
Configuration ↓	Maximum temperature at CRAC inlet (℃)at t=300s	Maximum temperature at CRAC outlet (℃) at t=300s
Baseline	34.1	15
Aisle of Active Tiles	32.4	18.3

Table 9 compares the maximum temperature of air entering and exiting the CRAC unit 300 s after failure for both scenarios. The CRAC inlet temperature at time t = 300s is lower for configuration with active tiles, than the baseline configuration for both failure scenarios. However, it is interesting to note the temperatures at the CRAC exit for B, the scenario of CRAC blower failure. For passive tiles, there is no increase in temperature from t= 0 s to 300 s. Because the blower fails, almost no air flows through the CRAC from the inlet, through the chilled water cooling coil, to the exit, and hence this temperature is that of the stagnant air at the inlet. However, for active tiles, the low pressure created by the running active tile fans actually draws some air through the CRAC internal system, and this air exchanges heat with the chilled-water loop, reducing the temperature from 32.4 °C to 18.3 °C.

Figure 26 shows the air flow rate through the CRAC unit over time for scenario B. The flow rate for both configurations drops to zero at t = 0 s, when the CRAC blowers fail, and then become stable at a constant value which is much lower for the baseline configuration compared to the active tiles configuration. For the baseline case the flowrate is not exactly zero as the server fans would be operational and would cause circulation of

small quantity of air. This supports the previous discussion of the trends in CRAC exit temperature for failure scenario (B).

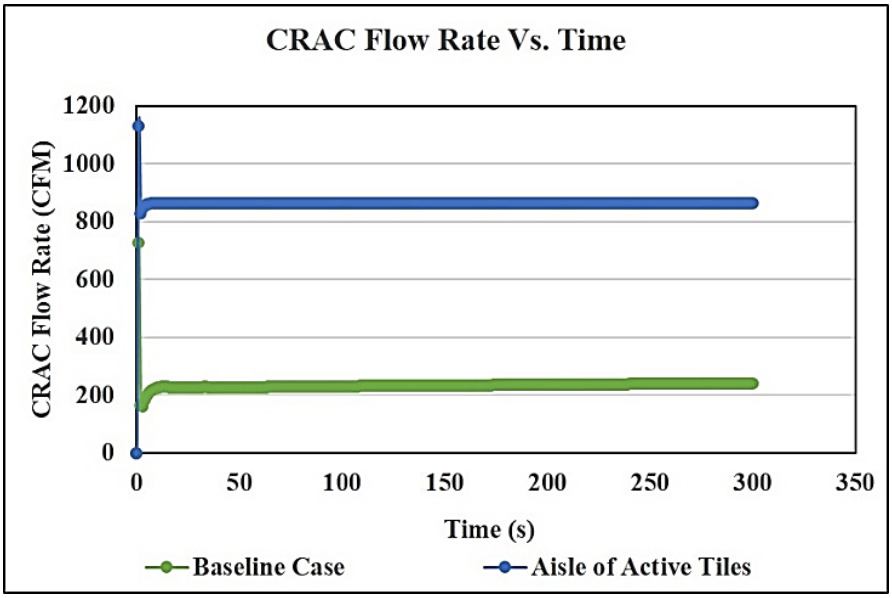


Figure 26 - CRAC Flow Rate as a Function of Time for Baseline Case and an Aisle of Active Tiles

As the air flow driven by the active tiles contributes directly to increasing ride-through time, this increase cannot be due to an increase in thermal mass alone. Thus, in either cooling failure scenario, active tiles increase ride-through time compared with passive tiles, and the increase is greater for the case of CRAC blower (*vs.* chilled water pump) failure.

2.4 Summary

An experimentally validated CFD model was developed using the commercial software package 6Sigma Room for three configurations of a data center room: a baseline case with all passive tiles, a configuration with a single active tile in cold aisle, and lastly a configuration with all active tiles in the cold aisle. The average overall discrepancy

between the numerical predictions and experimental measurements is found to be less than 4% for total tile flow rate and less than 1.7 °C for rack inlet temperature. Moreover, the CFD model captures, to a large degree, the qualitative trends in rack inlet temperature distribution. Parametric simulations conducted using his physics-based numerical models can be further used to generate training datasets for data driven modeling framework.

Furthermore, room-level CFD model using active tiles can be used to optimize the number and placement of active tiles to prevent hotspots without overcooling the entire data center. The results from selected case studies indicate that active tiles can have a significant effect on plenum pressure, which affects the total flow rate delivered to the cold aisle, as well as the percentage leakage from plenum to the room space. For a given CRAC blower speed, an increase in active tile fan speed increases the total tile air flow rate, and decreases the leakage of cold air in the plenum.

Transient simulations were run to investigate data center room response to CRAC failure scenarios when employing active, *vs.* passive, tiles. For failure of either the chilled-water pump or the CRAC blower, active tiles (compared with passive tiles) give a greater ride-through time for the IT equipment before critical temperatures are reached. For the case of CRAC blower failure, active tiles can maintain air circulation in the data center room, further lengthening ride-through time.

CHAPTER 3. ARTIFICIAL NEURAL NETWORK BASED PREDICTION OF TEMPERATURE AND FLOW PROFILE IN DATA CENTERS

In air-cooled data centers, predicting the temperature field and estimating and/or anticipating server cooling needs are both challenging due to the complex and non-isothermal air flows within the data center. Yet such accurate predictions are required to efficiently provision and distribute the cold air to ensure that each and every server functions within their temperature thresholds, while at the same time minimizing their energy consumption. Optimization of dynamic allocation of cooling resources in a data center requires model-based real-time thermal control [80], and this is presently unfeasible because we lack efficient and rapid modeling approaches. As mentioned in Chapter 1, the data driven modeling framework can potentially be employed to develop data center models with rapid predictive capabilities. The following section presents a detailed review of data driven compact models for thermal modeling of data centers.

3.1 Data Driven Modeling

Data Driven Modeling is based on analyzing the data about a system, in particular finding connections between the system state variables (input, internal and output variables) without explicit knowledge of the physical behavior [81]. Essentially, this approach represents a shift from “knowledge-based,” to “knowledge-learnt,” modeling.

Reference: J. Athavale, Y. Joshi, and M. Yoda, "Artificial Neural Network Based Prediction of Temperature and Flow Profile in Data Centers," in *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 871-880.(2018)

DDMs are particularly suitable for modeling data centers, given the non-linear nature of thermal transport, complexity of system operations and large number of associated metrics.

3.1.1 Proper Orthogonal Decomposition Based Modeling

Proper orthogonal decomposition (POD), a data driven statistical modeling framework, has been employed in a number of studies for rapid temperature prediction in data centers [82-87]. Also known as Karhunen-Loeve transform, POD involves expanding a set of data in terms of empirically determined basis functions for modal decomposition. Equation 3.1 describes the relationship between temperature field (*T*), POD coefficients (b_i) and POD modes (ψ_i) [83]

$$T = T_0 + \sum_{i=1}^{m} b_i \psi_i \tag{3.1}$$

where m is the number of retained POD modes in the decomposition.

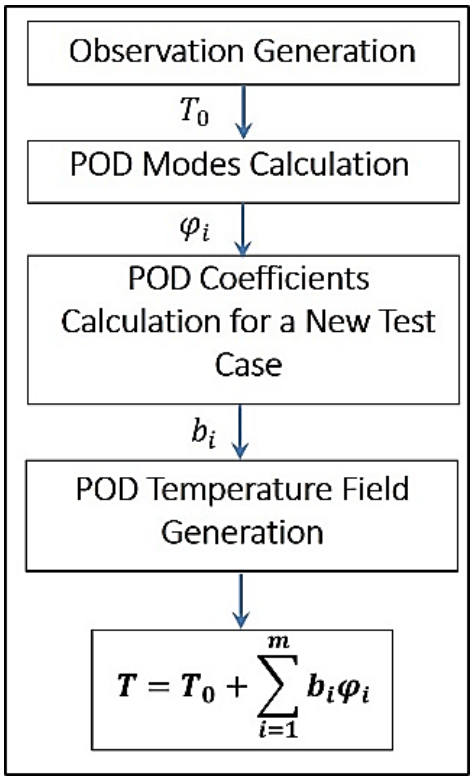


Figure 27 - POD Methodology

Figure 27 illustrates the POD framework. A large number of observations obtained from parametric CFD/HT simulations and/or detailed experimental measurements are used to determine the POD modes (ψ_i) of a thermal system. The key step wherein POD can be used to predict thermal profile as a function of system design variables is determination of POD coefficients (b_i) for a new test case, which is generally achieved by the Galerkin projection method, interpolation or flux matching process [82]. An important attribute of POD modeling that makes it well-suited for modeling temperature and airflow in data centers is that it captures more of the dominant dynamics within a given number of modes than any other linear decomposition [83, 85].

The scope of previous efforts employing POD for airflow and thermal modeling in data centers includes single parameter models, two parameter models, steady-state as well

as transient scenarios, and models developed from either experimental or computational data [82-87]. Samadiani et al., [83, 85] constructed a POD model with rack heat load and CRAC airflow rate as parameters for a small section of data center housing eight racks and one CRAC unit using results from CFD simulations. The simulation time reported was ~48 s and average prediction error for rack inlet temperature was found to be less than 1.5°C with maximum local errors as great as 2.5 °C. Ghosh et al., found that POD is accurate enough for interpolative prediction but has poor accuracy in extrapolations, *i.e.*, making predictions beyond the input parameter space [84]. Prediction using POD models do not require real-time measurements of pressure, air velocity or temperature, so POD models overcome one of the limitations posed by physics-based reduced-order models, as well as heuristic models. However, predictions for every new interrogation point in the input space require re-calculation of the corresponding POD coefficients, which is complex and time-consuming for an input parameter space of two or more dimensions.

As an extension to the traditional POD, Non-Linear Principal Component Analysis (NLPCA) methodology was used to model flow in a data center by Song et al. [88]. NLPCA is the nonlinear equivalent of standard PCA, and reduces the observed variables to a number of uncorrelated principal components. The method is used for data analysis and reduction and is capable of handling non-linear relationships between variables. Both POD and NLPCA models were developed using data (generated outcomes/results) from CFD simulations to predict velocity at three sensor locations for a data center with ten racks and one CRAC unit. The results indicated that the POD-based model could capture the flow field more accurately than the NLPCA model. It should however be noted that the previous

conclusion was based on comparison with CFD results for the same instances as the training data set, and hence is not a true indication of the prediction accuracy of the models.

3.1.2 Machine Learning Based Modeling

The second class of data driven models employ a machine learning/artificial intelligence framework for model development. Machine-learning models are ideal for systems which have multiple operational states based on interactions between the myriad electrical, mechanical and control parameters typical of data centers. From the perspective of IT equipment and resources, fuzzy logic-based models have been used to determine the relationship between the workload and resource needs of virtual containers and to guide resource allocation based on online measurements [89]. ANN models can be trained using either data from experimental measurements or CFD simulations. The ANN modeling framework can effectively transfer the computational complexity from model execution to model set-up and development, making it suitable for real-time prediction and optimization.

ANN models for thermal-aware scheduling and data center workload monitoring and analysis have been developed by Moore et al. [90-92]. The model developed was used to evaluate PUE sensitivity to data center operational parameters like server IT load, number of chillers running, etc., and to explore ways to improve energy efficiency in data centers. Shrivastava et al. [70] developed an ANN model to predict rack Capture Index (CI) using results from CFD simulations. Capture Index (CI) is a metric used to quantify rack cooling performance and is defined as the fraction of air ingested by a rack which originates from local cooling sources (perforated tile). The data center configuration under

consideration was that of a hard floor room with in-row coolers providing required cooling. The authors reported average root mean square errors of 3.8% and a maximum error of 27.4% for CI prediction. Finally, there have been a few studies on constructing neural network-based models capable of predicting complete temperature profile and tile flow rates (corresponding to a large number of neurons in the output layer) [93, 94]. Song et al. developed an ANN model to undertake a parametric analysis to study the effect of plenum characteristics on tile flow rate and temperature for a data center configuration with one cold aisle having 15 racks on each side and one CRAC unit [93]. An ANN-based model has also been constructed for prediction of PUE by Gao [95] using operational data collected over two years at one of Google's Data Centers.

3.2 Objective

The focus of this chapter is development of (ANN)-based (both steady-state and transient) models trained on datasets generated from offline CFD/HT simulations for rapid prediction of temperature and flow distributions in a data center (see Fig. 28). Using CFD simulation results to train ANN transfers computational complexity from model execution (in CFD) to model set-up and development. Multiple offline parametric simulations were conducted using the CFD/HT model developed in Chapter 2, and the corresponding results were then used to train the ANN model. The ANN model is then tested for its accuracy, and its potential utility in optimizing data center operations.

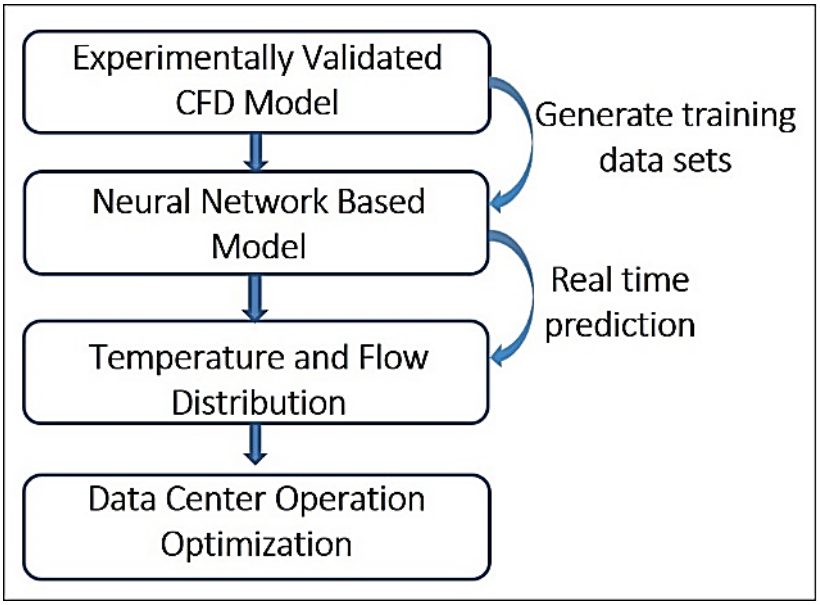


Figure 28-Proposed Framework

3.3 Artificial Neural Network (ANN) Modeling

ANN models are suitable for thermal modeling of data centers as they are capable of encoding the non-linear relationships between the input variables (here, CRAC return temperature $(T_{a,ret})$, flow rate (N_{CRAC}) and rack IT load $(\dot{Q}_{IT\ room})$ and output data (thermal and flow profiles), and produce outputs that fall in a continuous paradigm [96, 97].

Using CFD simulations for training the ANN transfers the computational complexity from model execution (as in CFD) to model set-up and development. Using a framework that requires a time-consuming one-time set-up (in training), but can then be rapidly executed multiple times, is required for real-time optimization and control. Another advantage of employing ANN is that the model can be later updated to a broader input parameter space once more data become available.

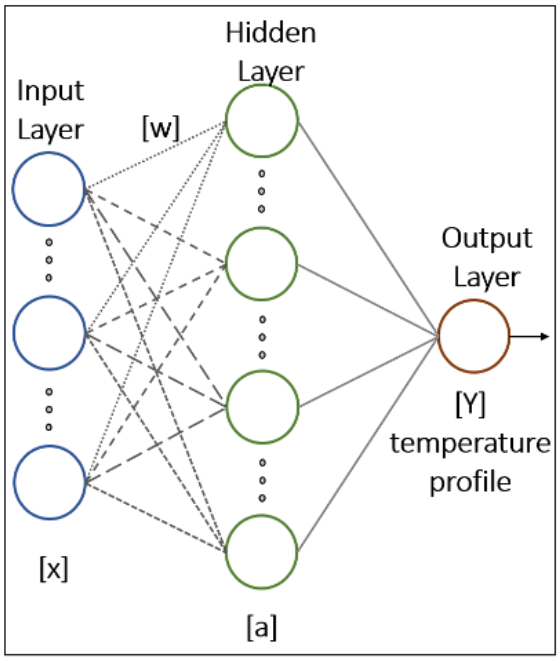


Figure 29 – Typical Neural Network Topology

The typical structure of a neural network, shown in Fig. 29, consists of an input layer, hidden layers and an output layer, each having multiple neurons. Each neuron in a given layer is linked to (*i.e.*, dependent on) every neuron in the preceding layer; this dependence is characterized by the weights of these links. Equation 3.2 represents the output of a neuron (y_j) in a layer where n_i is the number of neurons in the preceding layer, i is the index for neurons in the preceding layer, f is the non-linear activation function and b_{ANN} is the bias term associate with neuron under consideration.

$$y_j = f\left(\sum_{i=1}^{n} w_{ij} x_i + b_{ANN}\right) \tag{3.2}$$

Backpropagation [98] is the most widely used class of algorithms for calculating these weights. Weights are determined by minimizing an associated cost function, which in this case is given by equation 3.3 where *N* is the total number of temperature points.

$$Mean\ Square\ Error(MSE) = \frac{1}{N}\sum_{i=1}^{N}\left(T_{\Pr edicted} - T_{Actual}\right)^2 \tag{3.3}$$

3.4 Case Study I – Steady State Modeling:

A neural-network model was developed to predict steady-state rack inlet temperatures, tile flow rates, plenum pressure and CRAC supply temperature as a function of CRAC blower speed (N_{CRAC}), CRAC return air temperature set-point $(T_{a,ret})$ and IT load factor for row A and row B, LF_{rowA} and LF_{rowB}, respectively, for the data center room, where the IT load factor,

$$Load\ Factor = \frac{Actual\ IT\ Load}{Maximum\ IT\ Load} \tag{3.4}$$

3.4.1 Generating Training Data

Neural networks are an empirical modeling technique wherein the non-linear and complex relationship between the independent and the dependent (predicted) variables are "learned" by presenting a large number of training examples to the modeling framework. The multi-dimensional parameter space for the independent variable should be space-filling and non-collapsing [99]. Computational simulations are deterministic in nature and are not prone to the uncertainties inherent in experimental methods; hence, it is critical that the input parameter space is determined using a random sampling technique to avoid any

bias and introduce required variability in the training data for neural networks. Latin hypercube sampling (LHS) [100], a statistical method for generating a near-random sample of parameter values from a multidimensional distribution, ensures that the ensemble of random numbers are representative of the real variability. The LHS method was used to generate a random sample of 500 combinations of input parameters; corresponding parametric CFD simulations were used to generate the training dataset. Table 10 gives the independent variables (parameters) and their corresponding bounds and constraints.

Table 10- Input Parameter Space Definition for Neural Network Training Data

Independent Variable	Bound	Constraints
CRAC Blower Speed (N_{CRAC})	30%-100%	Constrained to be a multiple of 5 i.e. (30%, 35%, 40%,…100%)
CRAC Return Air Temperature set-point ($T_{a,ret}$)	18℃ to 30℃	Rounded off to be an integer value
IT load factor for row A (LF_{rowA})	0-1	Rounded off to first digit after the decimal point
IT load factor for row B LF_{rowB}	0-1	Rounded off to first digit after the decimal point

As mentioned previously, all the racks in the room have different capacities depending on the IT equipment and the number of occupied slots in the rack. Figure 12(b) (in Chapter 2) shows the maximum load capacity for each rack. Though all the racks in a given row have the same IT load factor, they have different absolute heat loads depending on their maximum load capacity. Table 11 gives the dependent (predicted) variables and fixed parameters for this study and Fig. 30 gives the location for the four temperature sensors in each rack.

Table 11-Predicted variables and Fixed Parameters for Neural Network Training Data

Dependent Variable	Description
Rack Inlet Temperature	4 temperature sensors per rack (Fig. 30); total **36 points**
Tile Flow Rate	1 for each perforated tile; total **10 points**
Plenum pressure	**1 point**
CRAC Parameters	Mean supply temperature, mean return temperature (measured) and cooling air flow; total **3 points**
Fixed Parameters:	Room layout and IT and cooling equipment configuration, tile porosity (see section B.2)

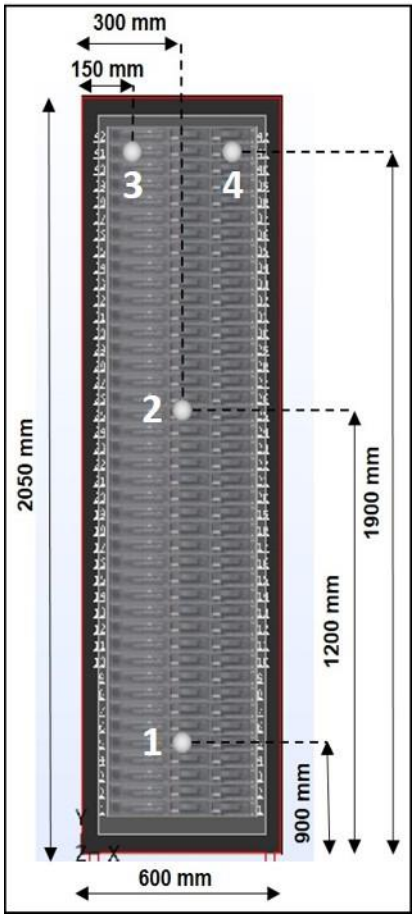
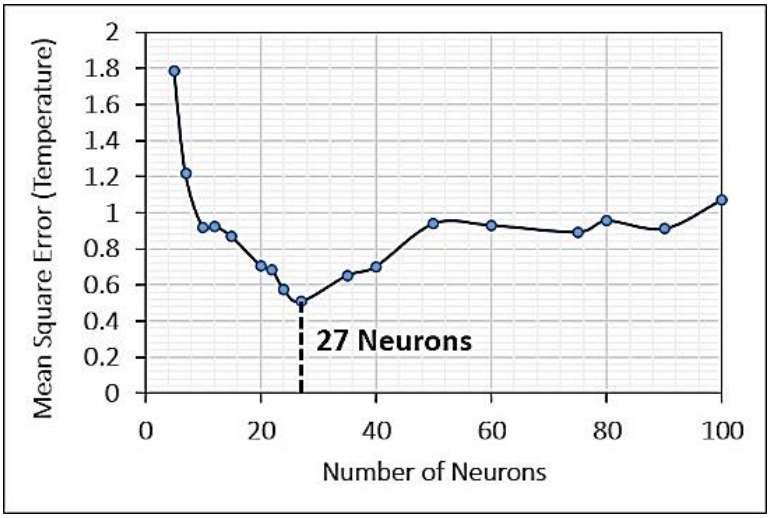


Figure 30 – Location of Temperature Sensors

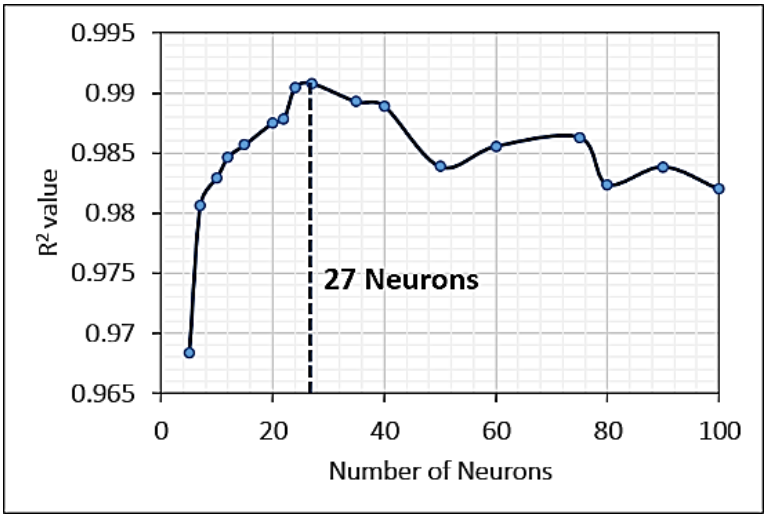
3.4.2 Model Selection

Two three-layer (12-27-36) and (12-13- 14) ANN consisting of one hidden layer, one input layer and one output layer were selected as the model architectures for rack inlet

temperature prediction and prediction of flow variables, respectively. Many empirical relations are available in the literature to determine the appropriate number of nodes for the hidden layer based on the number of number of nodes in the input and output layers [101-104]. Several configurations with varying number of nodes in the hidden layer were considered; the configuration which gave the lowest error (and therefore the highest R^2 value) for both training and testing data was chosen for further refinement (see Fig. 31).



(a)



(b)

Figure 31 - Comparison of Network Performance for Networks with varying Number of Neurons in Hidden Layer

The tan(gent) sigmoid function, given in equation 3.4, was the non-linear activation function for the hidden layer (Fig. 32) [105].

$$\tan sig(n) = \frac{2}{1+e^{-2n}} - 1 \tag{3.4}$$

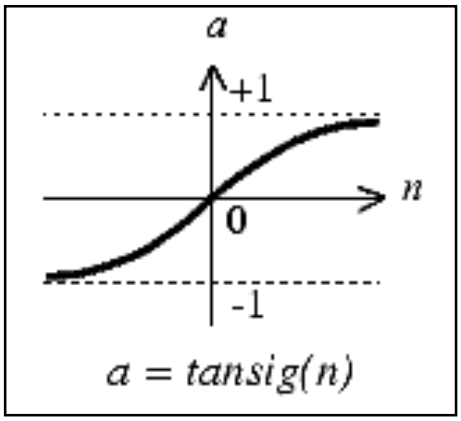


Figure 32 - Representation of Tangent Sigmoid Function

Given that the output of the ANN will be in the range [-1:1], because of the choice of activation function, the training dataset is first normalized such that it falls within the same range. A linear activation function was incorporated in the output layer.

While constructing data driven models, collection/generation of sufficient data to train a high-fidelity model is a time-intensive step. It is therefore desirable to determine an optimum size for the training dataset to minimize the time and computational resources required for model development without adversely affecting model accuracy. A number of ANN models were constructed by varying the number of samples used for training from 50 to 500. Figure 33 shows the training, validation and testing errors for the ANN models as a function of the number of training samples used for model development. The testing error decreases monotonically as the number of samples used for training increases until

~300 samples; for more than 300 samples, the decrease in the error is negligible. Hence, 300 samples were used to train the ANN model.

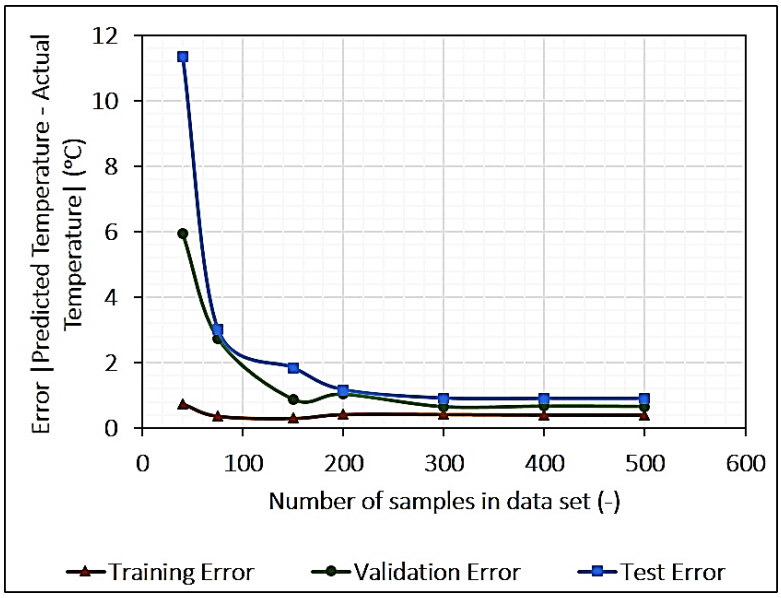


Figure 33 - Comparison of Network Performance for Networks with varying Number of Sample in Training Data Set

3.4.3 Model Training

Training the network involves determination of the weight coefficients that minimize the prediction error associated with the network. The training samples available were randomly divided into three sets—for training (70%), validating (15%) and testing (15%)—the network. The validation set helps to further refine the developed model, while the testing set helps to estimate prediction error bounds for the network developed.

The Levenberg-Marquardt backpropagation algorithm (LMA) [106], which provides a numerical framework to minimize non-linear functions, was used to train the network. LMA overcomes the flaws of both the gradient-descent (slow convergence) and the Gauss–Newton (prone to divergence) methods for neural-networks training by blending

the two algorithms and implementing each when appropriate during error minimization. The computational complexity associated with LMA is that of calculating a Jacobian matrix of partial derivatives of network error with respect to weights and biases, and is very efficient for training small- to medium-sized networks. LMA traverses the entire data set multiple times (epochs) minimizing the associated error and the training concludes when MSE is below a predefined threshold or if there is no further decrease in MSE for a certain number of consecutive epochs.

Figures 34 and 35 depict the training statistics for the ANN model developed here, with predicted values indicating ANN model-based predictions while actual values indicating results from CFD simulations used for training. Regression plots both for predicted rack inlet temperature and tile flow rate indicate that ANN model fits the data reasonably well with R^2 = 0.99, as shown in Fig. 34. There is greater variation around the perfect fit line (at an angle of 45°) for the predicted temperature values compared to that for the predicted tile flowrate values. This is expected because there is a more direct relationship between tile flow rate and input parameters (especially N_{CRAC}), which can be more easily captured than the indirect relation between rack inlet temperature and the input parameters. Error histograms for all the training instances for rack inlet temperature and tile flow rates (Figs. 35(a) and 35(b), respectively) show that most of the errors are centered around zero, further demonstrating that the ANN model is a good fit to the data.

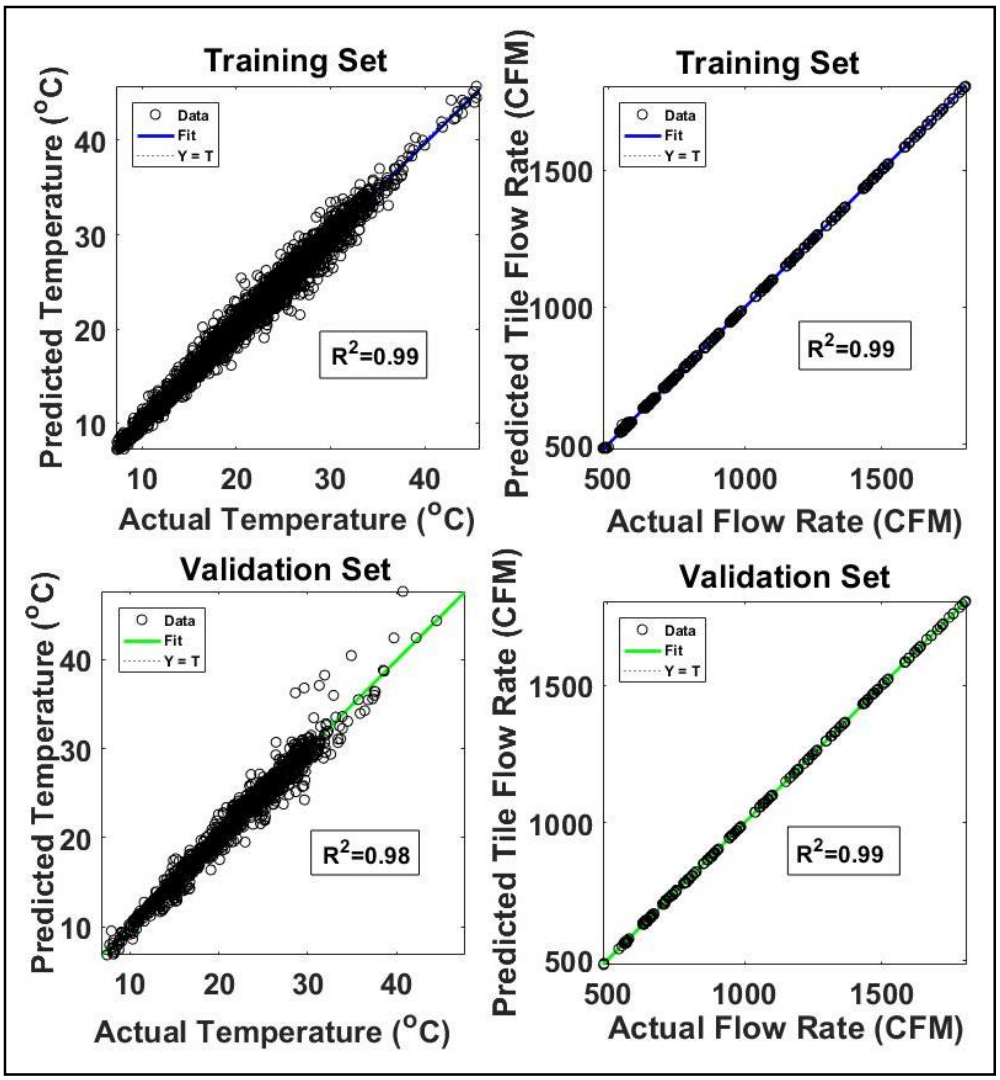
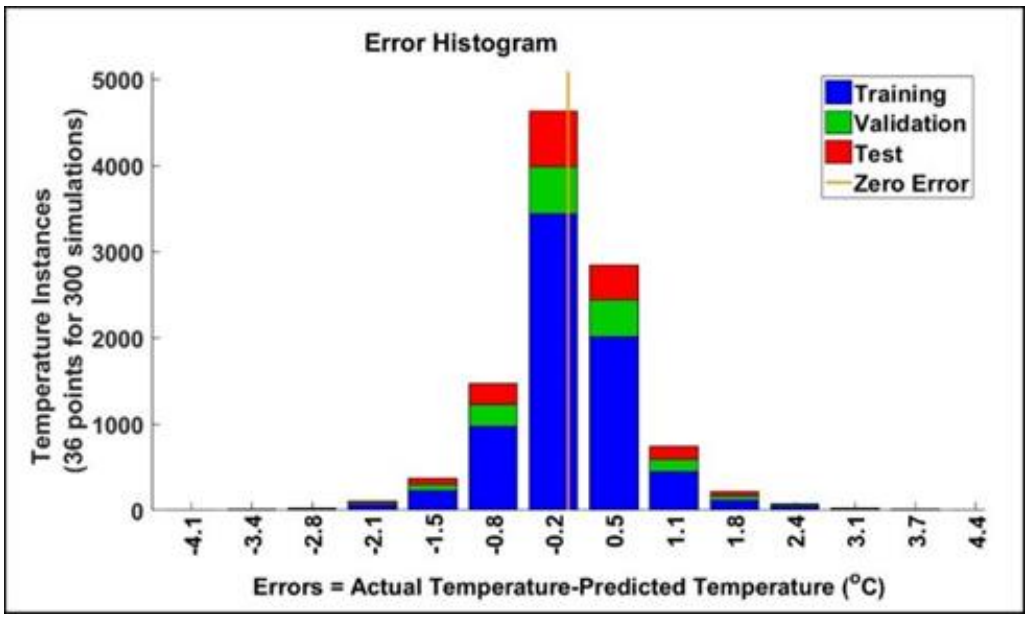
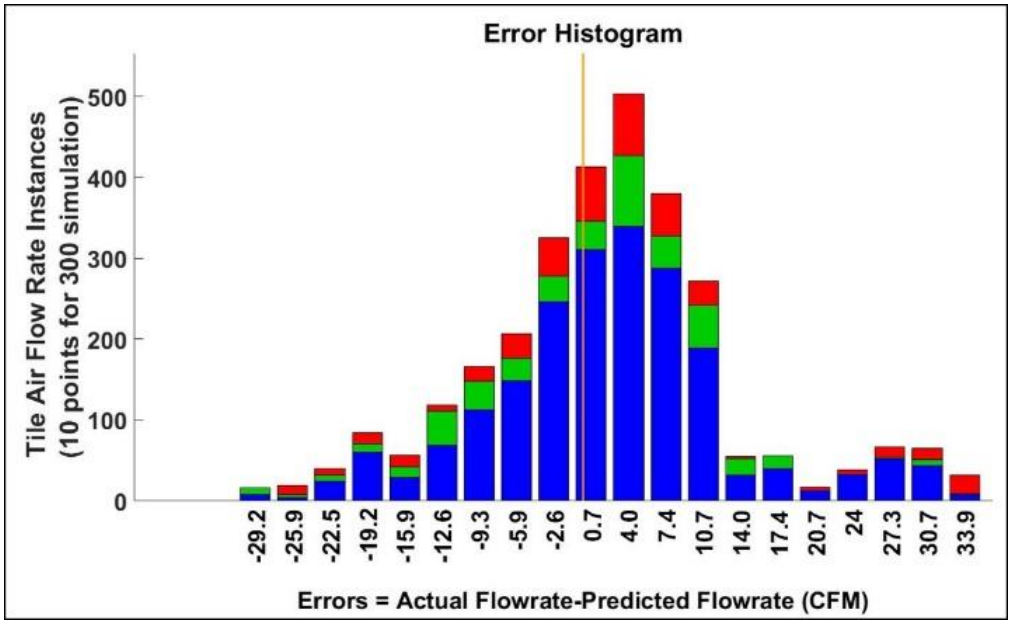


Figure 34 - Regression Plots for Predicted Vs. Actual Rack Inlet Temperature and Tile Flow Rate



(a)

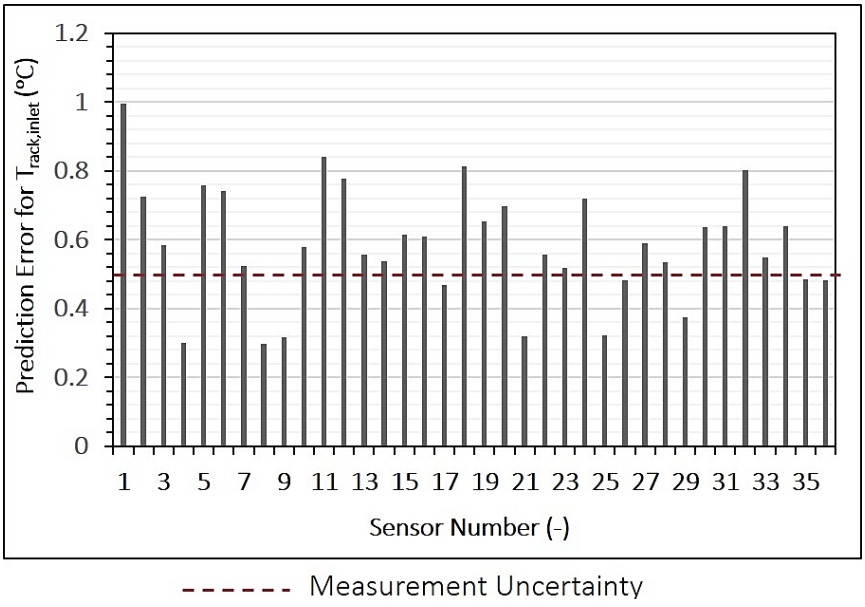


(b)

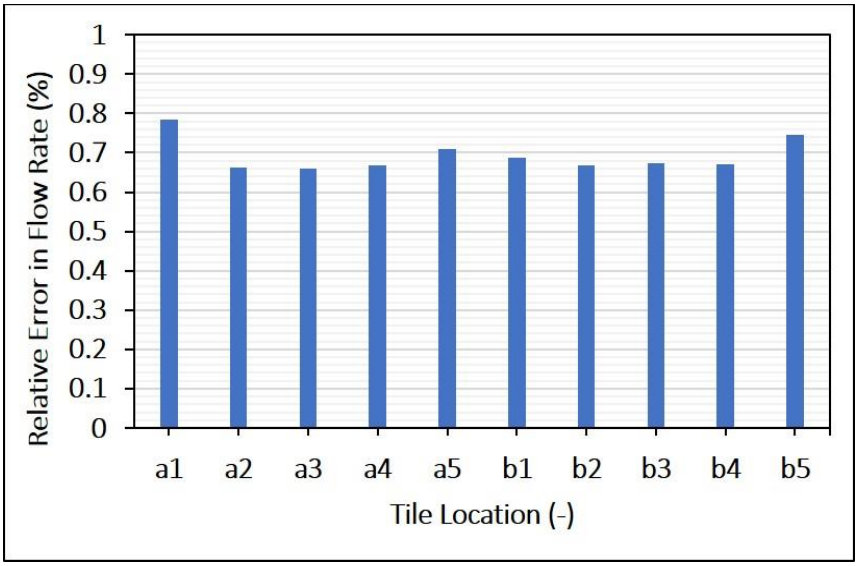
Figure 35 (a) - Error Histogram for Rack Inlet Temperature Prediction (b) Error Histogram for Tile Flow Rate Prediction

3.4.4 Model Testing

The predictive capabilities of this ANN are then tested using a fresh set of 33 CFD/HT simulation, which are derived from the same multi-dimensional input parameter space, and are hence interpolative in nature.



(a)



(b)

Figure 36(a) Prediction Error for Rack Inlet Temperature (b) Prediction Error for Tile Flow Rate

Figure 36(a) shows the prediction error associated with each of the 36 rack inlet temperature points averaged over 33 simulations. The average discrepancy between the predicted and actual temperatures for the 33 simulations is 0.6 °C, while the average relative error with respect to the actual temperature (given in °C) is 2.7%. Figure 36(b) gives the relative error in the predicted tile flowrates at each of the ten locations averaged over 33 simulations. It should be noted that all the error values in this case were below the experimental measurement uncertainty of ±5%. We therefore conclude that the ANN model developed is capable of predicting rack inlet temperature and flow variables with high accuracy.

3.5 Case Study II – Transient Modeling

A neural network model was also developed to predict temperature evolution in a data center for a transient scenario. The transient scenario considered here is a cooling failure scenario, where the chilled-water pump (CWP) fails while the CRAC blowers (CB) remain active to recirculate the air present in the data center room. Data centers are classified as mission-critical facilities and as such, it is important to ensure that their operating conditions are in accordance with ASHRAE recommendations at all times. Developing accurate predictive models for failure scenarios can provide valuable insights into temperature evolution and ride-through times in data centers to guide implementation of corrective measures and reduce the chances of IT equipment failure or damage due to overheating.

3.5.1 Generating Training Data

Figure 37 depicts the transient simulation used for training ANN model. Here, N_{CRAC}, $T_{a,ret}$ and $\dot{Q}_{IT\ room}$ are fixed parameters, while time *'t'* is the input parameter. The rack inlet temperatures at the 36 sensor points shown previously are monitored as the output variable in transient CFD simulation used for generating training data. As shown, the first 200 s of data (at 10 s intervals) were used for training the ANN and testing the interpolative predictive capability of the model, while the next 300 s of data were used to test for extrapolative predictive capability.

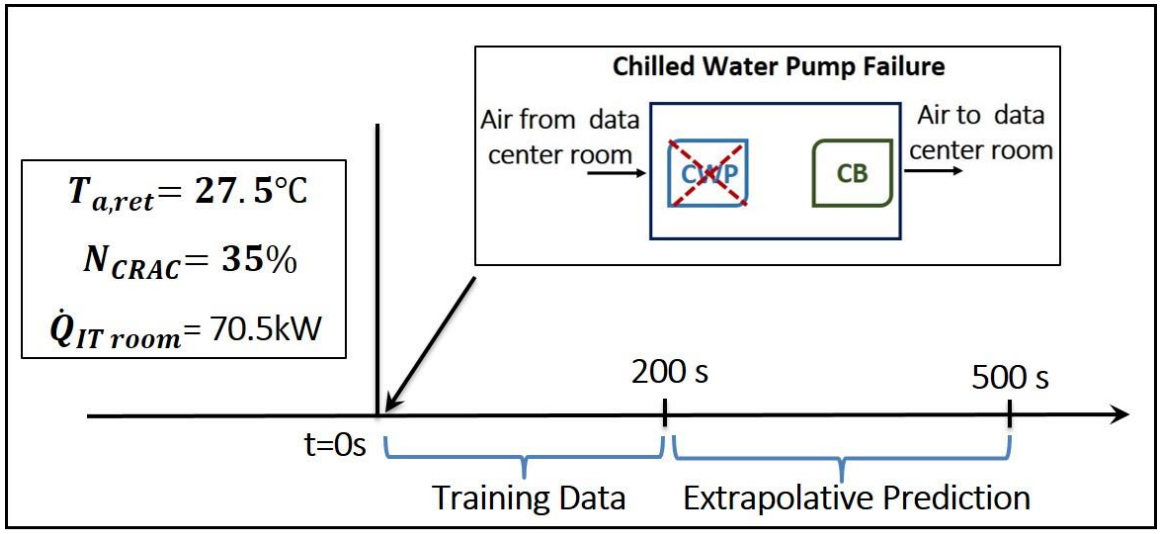


Figure 37 - Transient Scenario

3.5.2 Model Selection

A three-layer (12-27-36) ANN consisting of one hidden layer, one input layer and one output layer was selected as the model architecture for rack inlet temperature prediction.

3.5.3 POD Method

As mentioned before, Proper Orthogonal Decomposition (POD) has been used in a number of studies to rapidly model and predict data center parameters. Figure 27 illustrates the POD framework; further details can be found in [83].

A POD model for the same transient scenario was developed and results for the ANN and POD-based models were compared in terms of predictive accuracies and computational times. Figure 38 gives the energy captured by the POD modes; the POD model selected here retained all 20 POD modes. Energy captured by a particular POD mode is indicative of temperature variability explained by that mode.

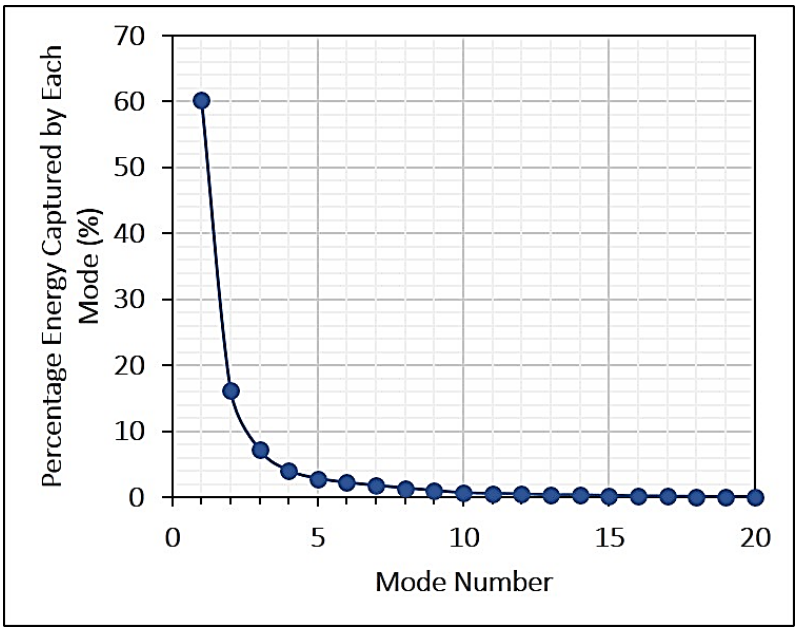


Figure 38 - Percentage Energy Captured with respect to Mode Number

3.5.4 Model Testing and Comparison

Figure 39 compares the absolute prediction error in the rack inlet temperature averaged over 20 interrogation instances (new prediction instances) for the ANN and POD

models. The interrogation instances are separate from the ones used for training but still within the 200 s interval over which the models are trained, and are therefore interpolative in nature. The prediction error for both modes is below measurement uncertainty in many cases, and < 0.9 °C in all cases. For most of the sensor points, the prediction error associated with ANN model is found to be lower than that associated with POD model. This indicates that the ANN model, which can capture non-linear relationships, can predict as well, if not better, than the POD model, which is the most optimal linear basis for a given problem.

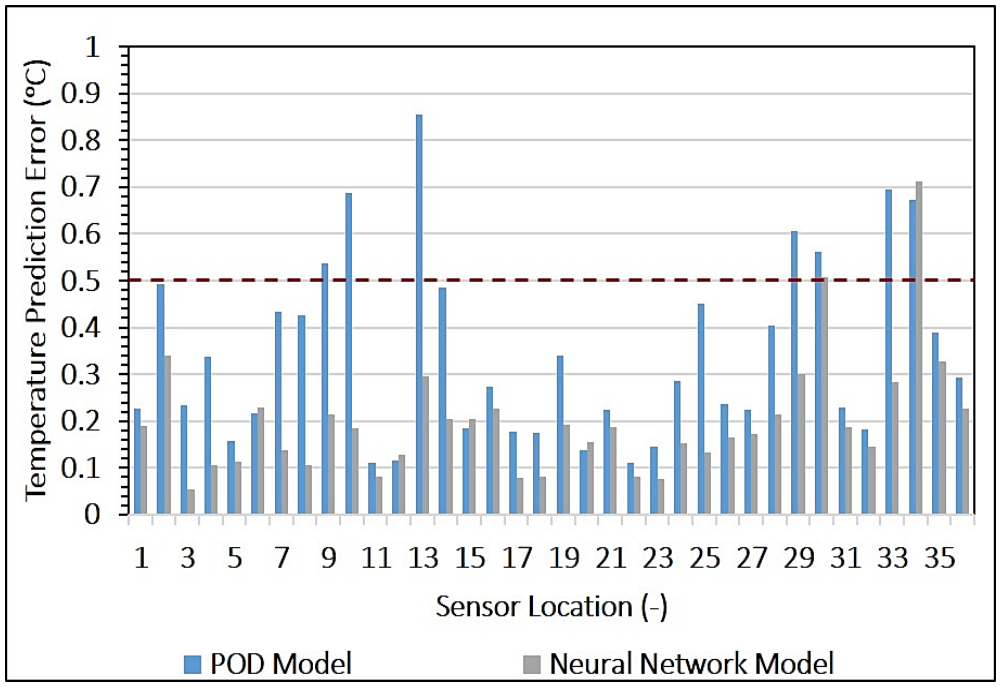


Figure 39 - Comparison of Interpolative Prediction Error in Rack Inlet Temperature for ANN and POD Model

The extrapolative accuracy for the two models was also compared here. Figure 40 gives the absolute prediction error averaged over all 36 sensor points over time. Note that the model is trained on data with 10 s intervals, while predictions are made every second. We suspect that the oscillations in the trend line are due to the mismatch in the training and prediction timesteps. Nevertheless, the graph shows that both models have temperature

prediction errors that increase monotonically with time progressing from the end of the training data. However, the rate of increase in the prediction error is much greater for the POD model.

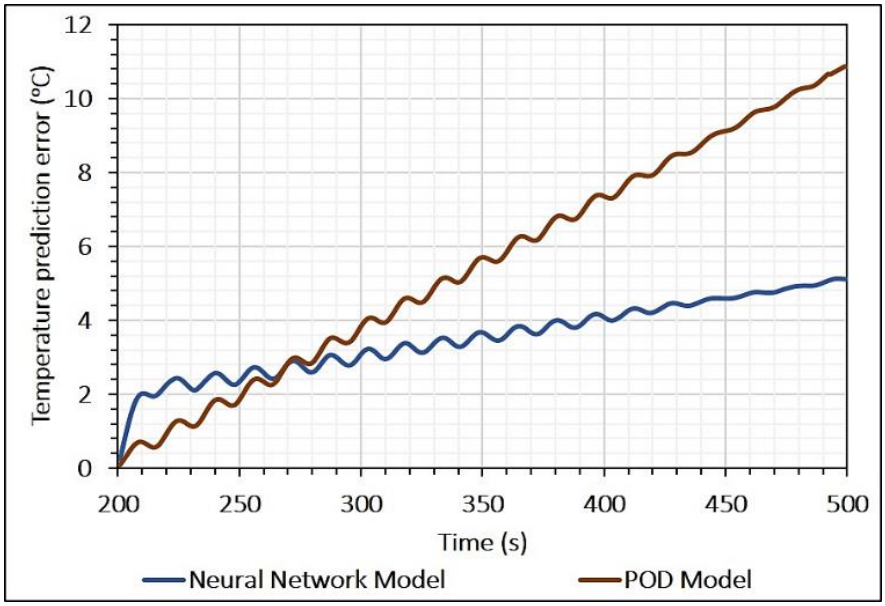


Figure 40 - Comparison of Extrapolative Prediction Error in Rack Inlet Temperature for ANN and POD Model

3.6 Comparison of Computational Time Requirements

Figure 41 compares the computational times required for temperature and flow profile prediction by the ANN model, the POD model and the room-level CFD model. The ANN model can predict rack inlet temperatures within seconds, which suggests that this model can be used for real-time control and/or optimization. It should be noted that this ANN model has a much lower spatial resolution compared with the room-level CFD model. Since the primary objective of this study was rapid and accurate prediction of rack inlet temperature points for optimization studies, however, this reduction in spatial resolution is acceptable for our purposes.

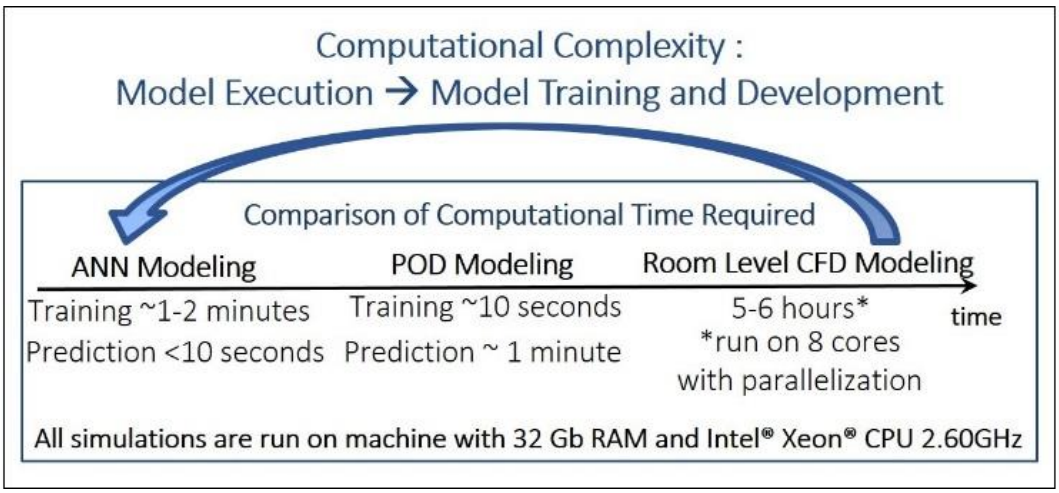


Figure 41 - Computational Times Required

3.7 Summary

A rapid ANN-based model for predicting temperature and velocity profiles in data centers was developed. The ANN model was trained using 300 experimentally validated CFD simulations using the CFD software Future Facilities 6Sigma Room. The resultant steady-state model reasonably accurate within generally applicable operational bounds typical in data centers for CRAC blower speed (based on equipment specifications), return air temperature set-point (based on ASHRAE specifications) and non-uniform IT load distribution. Verification tests suggest that there is good agreement between the ANN model and CFD simulations with an average error of < 0.6°C for rack inlet temperature prediction and 0.7 % for tile flow rate prediction.

An ANN model was also developed for a transient scenario; this model was tested and compared to a POD model developed using the same training data. Although the model seems to have low interpolative prediction errors, the extrapolative prediction errors are quite high, and appear to be directly proportional to the (here, temporal) distance of the interrogation point from the input parameter space.

The computational time required for prediction for a new set of input parameters is of the order of seconds for the ANN model, *vs.* hours for the CFD model. The rapid and accurate predictive capabilities of this ANN model suggest that such models can be used for real-time optimization and control.

CHAPTER 4. COMPARISON OF DATA DRIVEN MODELING APPPOACHES FOR TEMPERATURE PREDICTION IN DATA CENTERS

The cooling demands of IT equipment in data centers vary both temporally and spatially due to the unsteadiness of the workload and its random spatial distribution. This IT workload profile, coupled with lack of dynamic provisioning of cooling resources, leads to conservative set-points for the cooling infrastructure based on peak IT load demands. Various frameworks for dynamic allocation of cooling resources are being explored to reliably handle time-varying IT loads while minimizing cooling energy consumption. Implementation of real-time control makes it possible to use higher temperature set-points without the risk of server over-heating, and can therefore reduce energy consumption. As mentioned in Chapter 3, a key requirement for such a framework is the capability to rapidly and accurately predict the temperature field to efficiently provision and distribute the cold air while meeting server inlet air temperature limits.

This study describes the development and implementation of two different DDMs, namely Support Vector Regression (SVR) and Gaussian Process Regression (GPR), for predicting air temperature fields in data centers. SVR and GPR are kernel-based machine learning techniques that, like ANN, are capable of capturing non-linear relationships between inputs and outputs. To the best of our knowledge, this is the first application of these techniques to thermal modeling and temperature prediction in data centers. This

Reference: Athavale, J. ,Yoda, M., and Joshi, Y., "Comparison of Data Driven Modeling for Temperature Prediction in Data Centers", International Journal of Heat and Mass Transfer, (submitted 2018)

study also compares the performance of all the four DDMs (ANN, SVR, GPR and POD) developed, to provide guidance on selecting an appropriate data driven modeling approach.

It should be noted that the data center room configuration is as in Figs. 12(a) and 12(b) (in Chapter 2), and the dataset used for training the steady state models was generated using method described in section 3.4.1 (Chapter 3).

4.1 Description of Data Driven Modeling Methods Tested

Data driven modeling is a modeling paradigm which enables learning from a set of observations. It is especially suitable for modeling data centers characterized by non-linear thermal transport, complex system operations and large number of associated metrics. The following section gives a brief description of two (SVR and GPR) of the four data driven modeling techniques used in this study. ANN and POD based modeling frameworks have been detailed in Chapter 3. Table 12 provides a qualitative comparison of some of the distinctive features of the four techniques. Equations 4.1 and 4.2 give the general forms of the steady-state and transient models, respectively, obtained using the four data driven models. Here, the location co-ordinates of the temperature points being monitored (x, y, z) are constant.

$$T_{rack\ inlet} = f(N_{CRAC}, T_{a,ret}, LF_A, LF_B \mid x, y, z) \tag{4.1}$$

$$T_{rack\ inlet} = f(t \mid N_{CRAC}, T_{a,ret}, LF_A, LF_B, x, y, z) \tag{4.2}$$

4.1.1 Support Vector Regression (SVR) Modeling

SVR analysis is a machine learning tool first developed by Vapnik in 1992 [107]. SVR is capable of encoding non-linear relationships between input-output pairs in a given input data set by mapping the data in a high dimensional feature space, where the number of dimensions corresponds to the number of inputs, and performing linear regression in that space. The mapping function employed is called the "kernel function". Consider a set of data (x_n, y_n) where x_n is the vector of independent variables; y_n is the value of the dependent variable and the integer n = 1,2, ... N, where N is the total number of data pairs. The goal of SVR is to find a function, $g(x)$, such that $g(x_n) = y_n$ .The function $g(x)$ can be written as [108]:

$$g(x) = \sum_{n=1}^{N} w_n \varphi_n(x) + b \tag{4.3}$$

where, $\{\varphi_n(x)\}_{n=1}^{N}$ is a function representing the input parameters, and w_n b are coefficients determined by minimizing the following loss function:

$$P(g(x)) = \frac{C}{N} \sum_{n=1}^{N} E_{\varepsilon}(y_n, g(x_n)) + \frac{\|w\|^2}{2} \tag{4.4}$$

Here, $\|w\|^2 / 2$ is the flatness term, and C is the error penalty or regularization parameter, which determines the trade-off between training error and model flatness (simplicity).

Different loss (or risk) functions can be formulated for different applications, each with a different error € value. One of the most common loss functions for SVR, Vapnik's ε-insensitive loss function

$$E_{\varepsilon}(y_n, g(x_n)) = \begin{cases} 0 & \text{if } |y_n - g(x_n)| \le \varepsilon \\ |y_n - g(x_n)| \le \varepsilon & \text{otherwise} \end{cases} \tag{4.5}$$

was used here, where ε represents the error value that can be tolerated by the model. It can be shown that the regression function can also be represented as follows [108]:

$$g(x, \alpha, \alpha^*) = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*) K(x, x_n) + b \tag{4.6}$$

where α_n and α_n^* are Lagrangian multipliers such that $\alpha_n \alpha_n^* = 0$ and $\alpha_n, \alpha_n^* \ge 0$ $K(x, x_n)$ is the kernel (mapping) function mentioned previously. Commonly used kernel functions in SVR include linear and polynomial (quadratic, cubic), the radial basis, and the sigmoid and Laplacian. The functional form of the kernel determines which vectors in the training set most strongly influence the regression and the form of the estimator [109, 110]. The radial basis kernel function

$$k(x, x') = \exp\left[\frac{-\| x - x' \|^2}{2\sigma^2}\right] \tag{4.7}$$

was used in this study where $\| x - x' \|^2$ represents the square of the Euclidian distance between the feature vectors and σ is a free parameter. For model development, 75% of the samples in the dataset were used for training while the remaining 25% were used to

perform hold-out validation according to the criterion given in [111]. Hold-out validation involves setting aside a subset (here, 25%) of the available data and using this subset to test the performance of the developed model. In terms of application, it is important to note that SVR in its standard form is designed to handle datasets with a single output value, *vs.* an output vector (*e.g*. 36 rack inlet temperature points for a complete temperature profile) corresponding to a given input parameter vector. Thus, to obtain the entire temperature profile, 36 SVR models corresponding to individual temperature points were trained, then combined. This increases the amount of effort to training the SVR model compared with ANN- or POD-based models. The SVR model in this study was developed using standard functions in Matlab R2018a commercial package. For a more extensive description of SVR technique, refer to [112].

4.1.2 Gaussian Process Regression (GPR) Modeling

GPR is a kernel-based machine learning technique for non-linear regression problems, similar to SVR. A Gaussian Process (GP) is a set of random variables, such that any finite subset of these variables has a joint Gaussian distribution. If $\{f(x), x \in R^d\}$ is a GP, then given n observations $(x_1, x_2, \cdots x_n)$, the joint distribution of the random variables $(f(x_1), f(x_2), \cdots f(x_n))$ is also Gaussian. As a distribution, a GP is characterized by a mean function $m(x)$ and covariance function, $k(x, x')$ [113]. Due to the Bayesian context of its formulation and interpretation, GPR is probabilistic in nature, and gives prediction intervals, instead of a specific prediction point, for new interrogation parameters and thus inherently accounts for modeling uncertainty when used for interpolation and prediction [114].

Given a set of input-output parameters (x_n, y_n), a standard linear regression model with Gaussian noise can be represented as [115]:

$$g(x) = x^T w_{gpr} \tag{4.8}$$

$$y = g(x) + \varepsilon_{gpr} \tag{4.9}$$

where x is the input vector, w_{gpr} is a vector of weights of the linear model, g is the function value and y is the observed target value. It is assumed that observed target values y differ from the function values $g(x)$ by additive noise (see equation 4.9), which is an independent Gaussian distribution with zero mean and variance σ_n^2:

$$\varepsilon_{gpr} \sim N(0, \sigma_n^2) \tag{4.10}$$

For a more general case of non-linear regression,

$$g(x) = f(x) + h(x)^T \beta \qquad where,\ f(x) = GP(0, k(x, x')) \tag{4.11}$$

where $f(x)$ is a zero mean GP with covariance matrix k. Here, $h(x)$ are fixed basis (kernel) functions that transform the original feature vector x to a higher dimensional parameter space (as in SVR), and β is the vector of basis function coefficients.

The most common covariance function, the Gaussian squared-exponential (equation 4.12):

$$k(x, x') = \sigma_f^2 \exp\left[\frac{-(x - x')^2}{2l^2}\right] \tag{4.12}$$

was also implemented in this study. Here, l is the length parameter and σ_f is the function variance.

Training a GPR model entails determination of basis function coefficients (β),the free parameters of the covariance function (referred to together as hyperparameters $\theta = \{\sigma_f, l\}$), and the error noise variance (σ_n). Like SVR, GPR cannot accommodate multi-output predictions, and thus a separate model for each rack inlet temperature point prediction was developed. The training dataset was divided into two parts, with 75% of the data used again for training, while the remaining 25% was employed for hold-out validation. The GPR model in this study was developed using standard functions in Matlab R2018a commercial package. Further details regarding the GPR approach can be found in [115].

Table 12 – Summary of DDM Methods Tested

Feature	ANN	SVR	GPR	POD
Model Classification	Machine learning based technique	Machine learning based technique	Machine learning based technique	Statistical technique
Model Capability	Non-linear modeling technique	Non-linear modeling technique	Non-linear modeling technique	Linear modeling technique
Input Parameter Space	Suitable for modeling systems with multi-dimensional input parameter space	Suitable for modeling systems with multi-dimensional input parameter space	Suitable for modeling systems with multi-dimensional input parameter space	Suitable for modeling systems with 1D or 2D input parameter space
Output prediction Space	Typically can handle multiple output points (Ex: all temperature points in a room) in a single model	Typically cannot* handle multiple output points (Ex: all temperature points in a room) in a single model	Typically cannot* handle multiple output points (Ex: all temperature points in a room) in a single model	Typically can handle multiple output points (Ex: all temperature points in a room) in a single model
	Non-kernel based method	Kernel based method**	Kernel based method**	Non-kernel based method
Uncertainty accountability	Deterministic	Deterministic	Stochastic	Deterministic
* For SVR and GPR separate model for each temperature point was created and combined to obtain complete temperature profile				
** Kernel based methods require some prior knowledge about the system and input-output relationship				

4.2 Results and Discussion

The following sections titles Steady State Modeling as well as Transient Modeling present a comparative analysis of results for all the models developed in this study. It should be noted that ANN, SVR and GPR models were used to model the steady state, with a multi-dimensional input parameter space and the transient scenario, while POD was employed only for the transient scenario, with a 1D input parameter space).

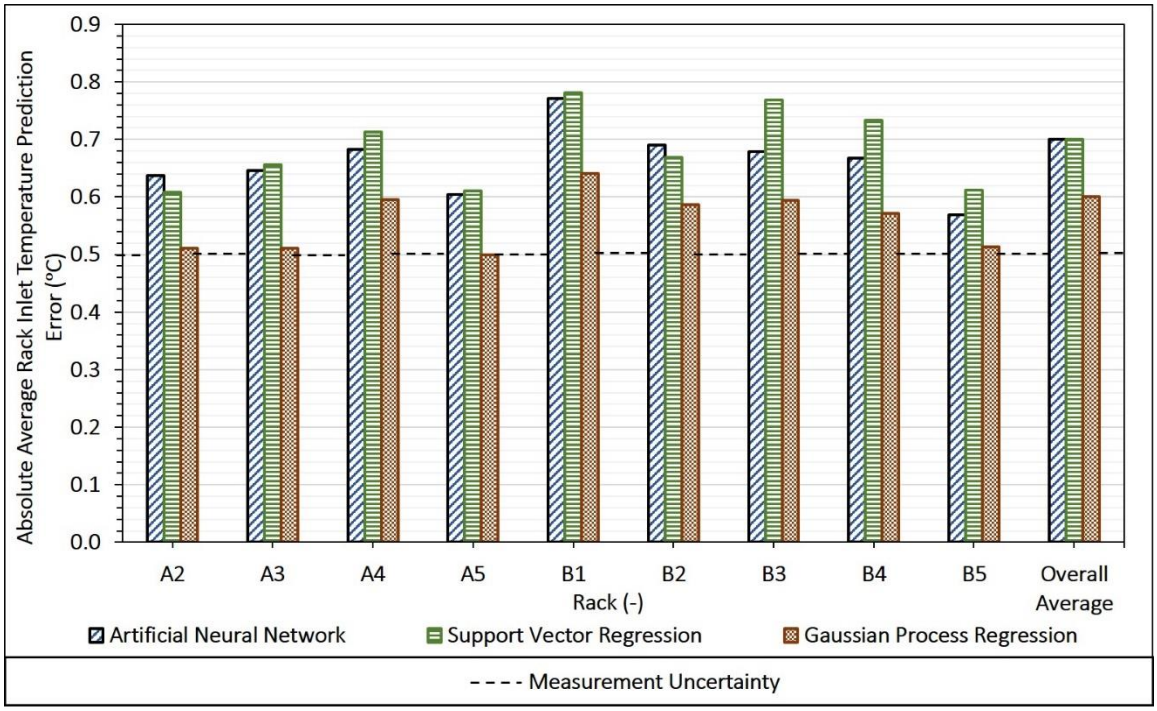
4.2.1 Steady State Modeling

A test dataset was generated using 30 new simulations to evaluate the accuracy of the three steady-state models. The values of the input parameters for the test dataset were distinct from those of the training dataset but were obtained from the same multi-dimensional input parameter space. Hence, the results below are representative of the interpolative prediction error in rack inlet temperature.

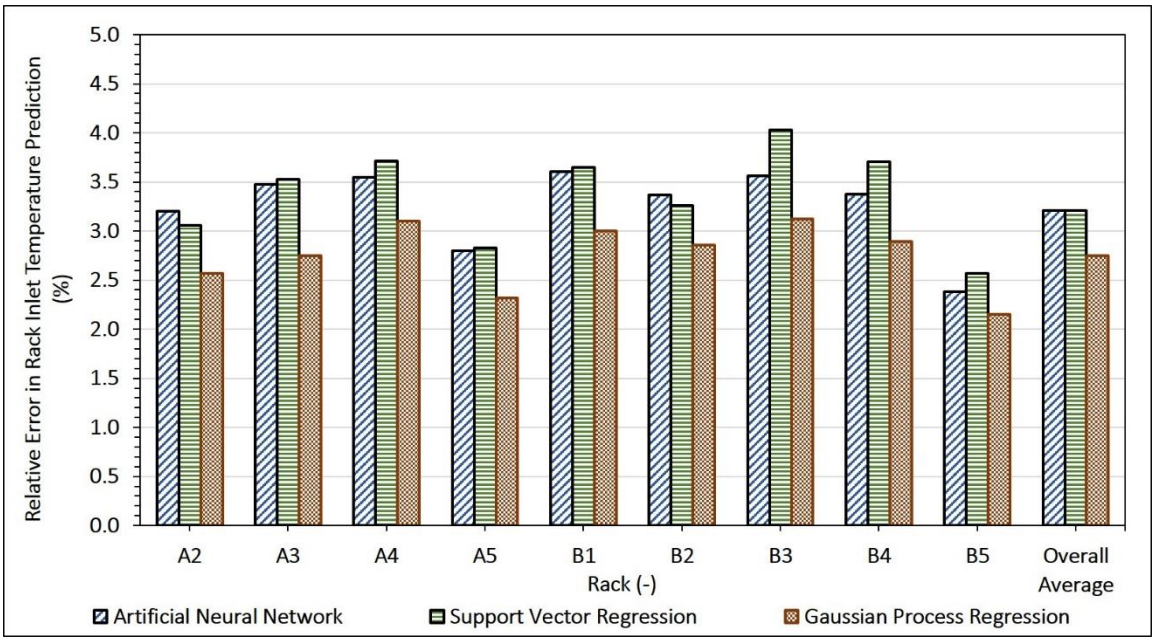
4.2.1.1 Interpolative Accuracy

Figures 42(a) and 42(b) depict the absolute and relative prediction errors in rack inlet temperature, respectively, for the ANN, SVR and GPR models averaged over 30 test simulations and for each of the nine racks (averaged over the four temperature points associated with each rack). The relative error is obtained by normalizing the absolute error with respect to the corresponding actual rack inlet temperature in °C.

The overall average discrepancy in predicted rack inlet temperatures is also shown, and found to be smallest, 0.6 °C (2.7%), for the GPR model; this error is 0.7 °C (3.2%) for both the ANN and SVR models. It should be noted that the prediction error is slightly greater than the actual measurement uncertainty in the temperatures of ±0.5 °C. We therefore conclude that these three data driven models can predict rack inlet temperatures with good accuracy.



(a)



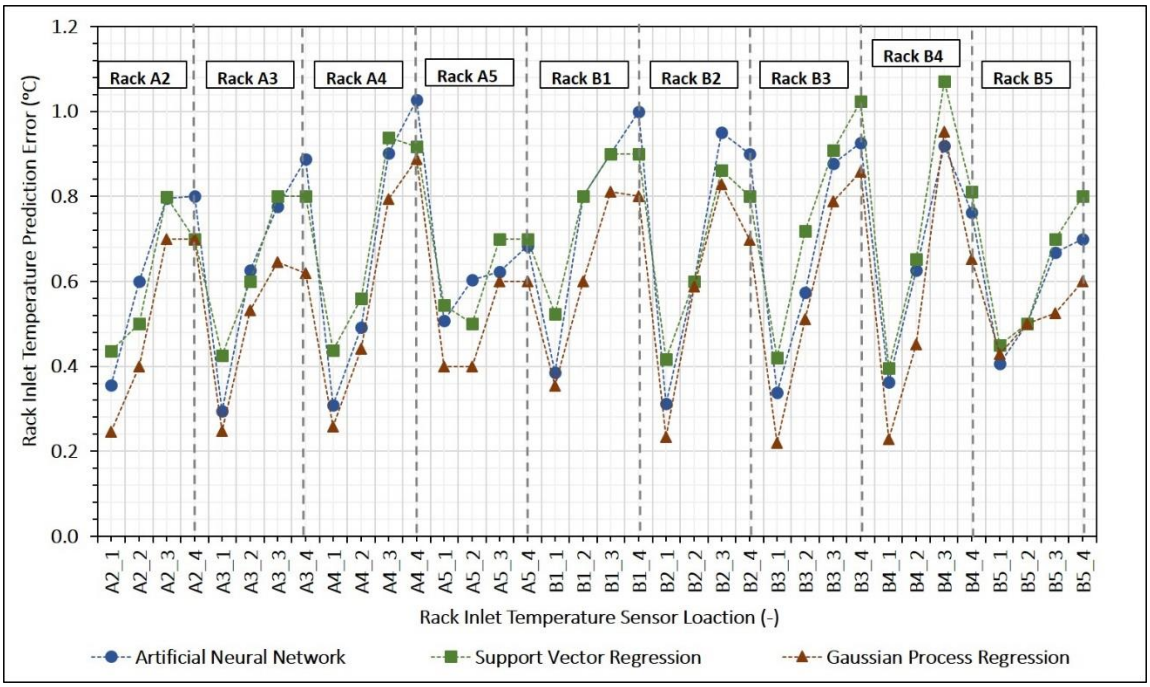
(b)

Figure 42(a) - Absolute Prediction Error for Rack Inlet Temperature (b) Percentage Relative Error for Rack Inlet Temperature

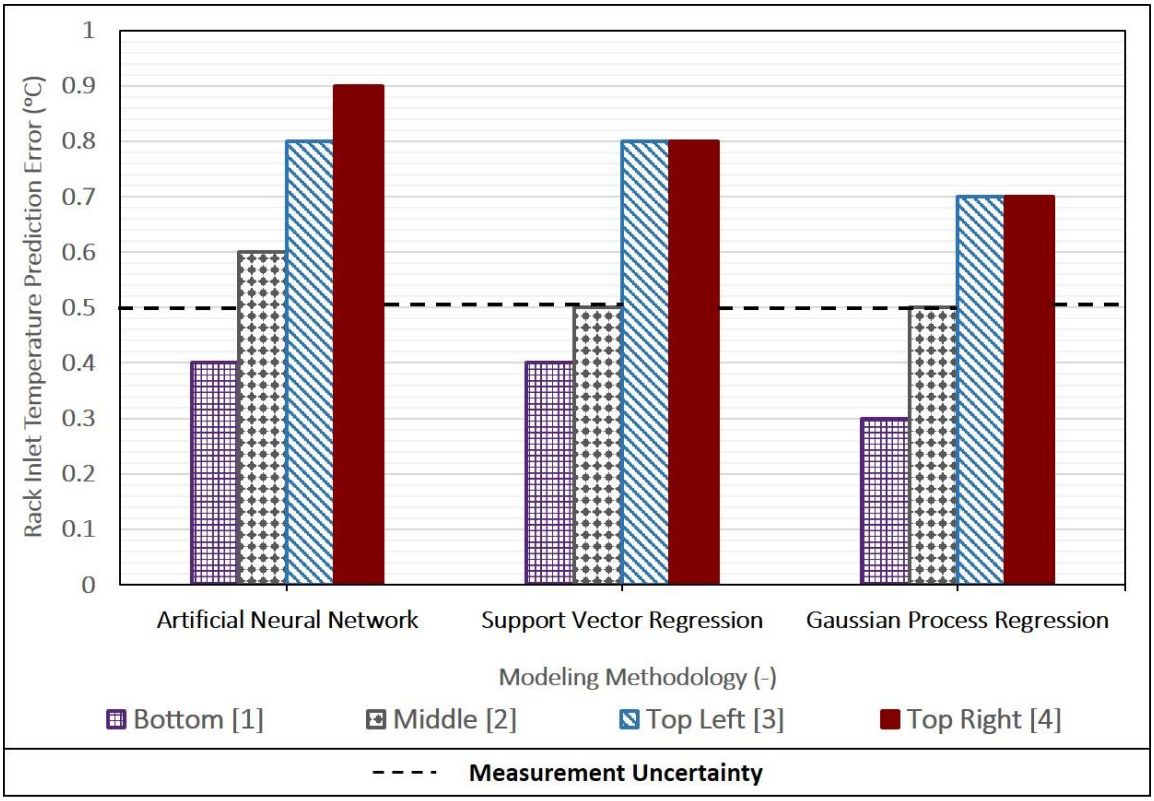
4.2.1.2 Relationship between Prediction Error and Temperature Measurement Location

Figure 43(a) presents a more detailed distribution of the discrepancy in temperature predictions for every temperature sensor location. The plot shows that this prediction error increases with height from the perforated floor for all three models.

Figure 43(b) shows instead the prediction error averaged over similar temperature sensor locations for these models. The prediction error at the top two sensors (3 and 4) is approximately twice that of the lower two sensors for all models. This larger error could be due to air recirculation effects, which are not accurately captured by the data driven models (and, in many cases, by CFD/HT models as well). Recirculation is an undesirable flow pattern that occurs in data centers due to a mismatch between the amount of cooling air supplied through a perforated tile at a given location, and that required by a rack at that location (see Fig. 44). Server fans at the top section of the rack can pull in hot air from behind the racks to compensate for the deficit in cooling air, leading to hot spots. The extent of recirculation in underfloor supply air-cooled data centers depends on many parameters including location, computational load for given racks, server fan speeds, CRAC blower speed, type of perforated tiles, etc., and is difficult to estimate with even CFD/HT models. As recirculation mainly occurs near the top of the racks, the inability to capture these effects affects the accuracy of the data driven model predictions in those specific locations.



(a)



(b)

Figure 43 (a) - Rack Inlet Prediction Error for Individual Sensor Locations for Every Rack (b) - Rack Inlet Temperature Prediction Error with respect to Sensor Location

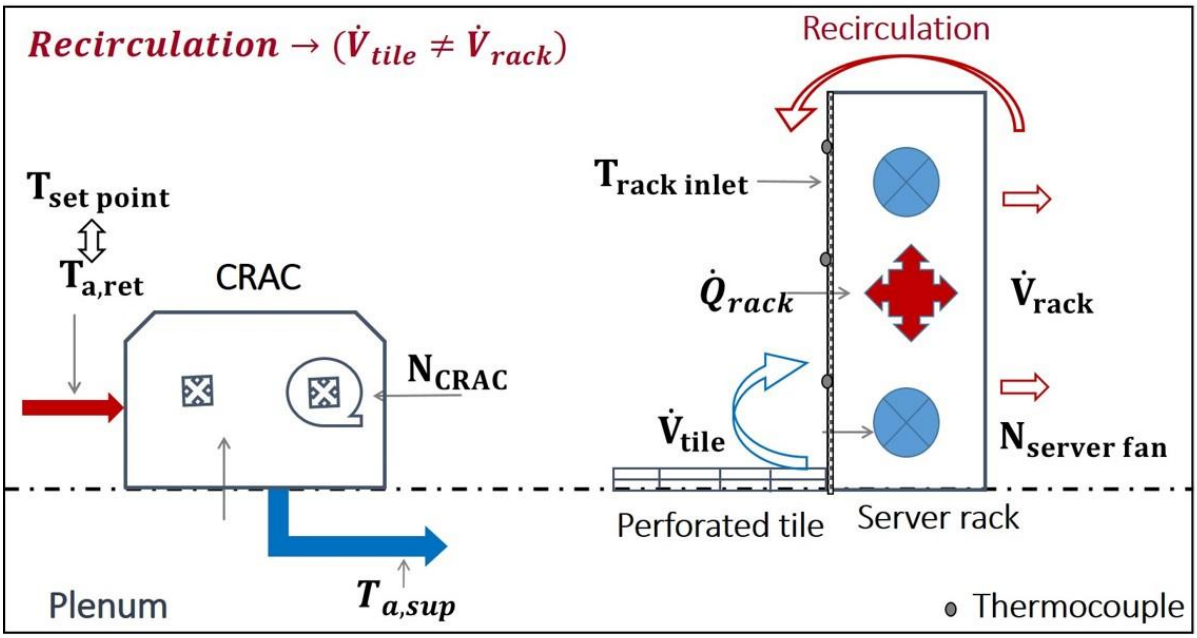


Figure 44 - Flow Path and Recirculation in Data Centers.

4.2.1.3 Number of Samples Required

When constructing data driven models, collecting/generating sufficient data to train a high-fidelity model is usually the most time-intensive step. It is therefore useful to determine the optimum size of the training dataset to minimize the time and computational resources required to develop the model without compromising the accuracy of the model. Moreover, there are many instances where only a limited amount of data is available, for example when modeling newly commissioned/built data centers. In such cases, it is useful to determine how the prediction accuracy depends on as the size of the training dataset for different data driven modeling frameworks, and could guide selection of a suitable modeling framework and give an estimate of the expected error.

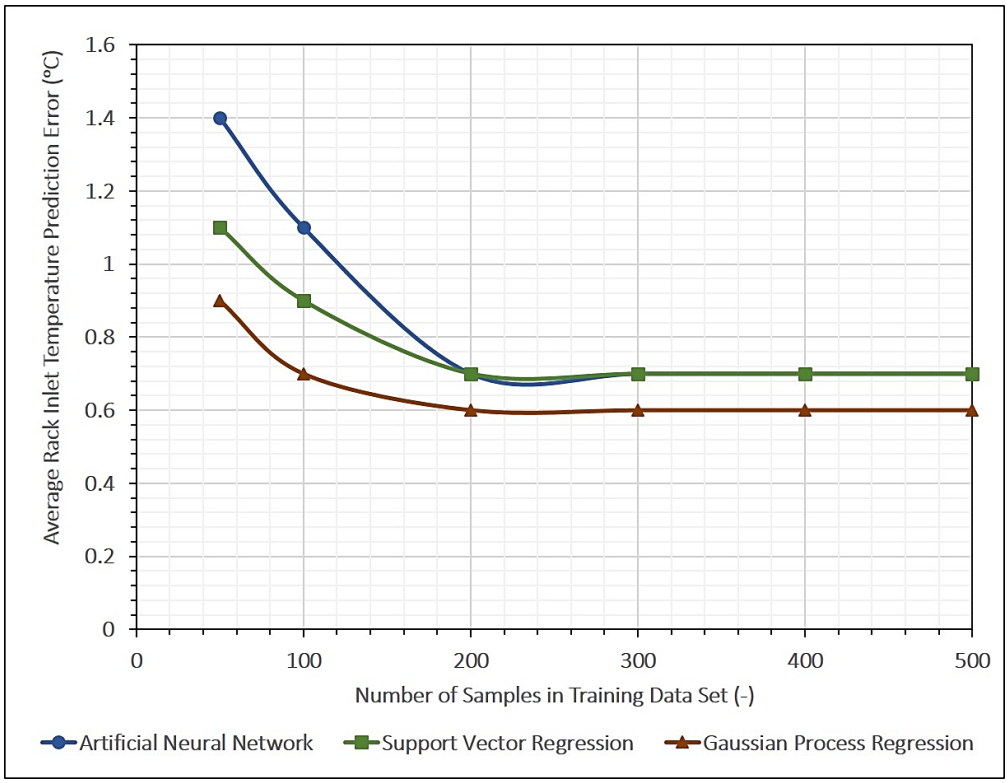


Figure 45 - Comparison of Modeling Framework Performance for Models with varying Number of Samples in Training Dataset

Several models were constructed by varying the number of samples used for training from 50 to 500 using all three modeling frameworks (ANN, SVR and GPR). Figure 45 shows the error in predicted temperature for ANN, SVR and GPR models as a function of the number of training samples used to develop the model. There is a monotonic decrease in the temperature error with increasing number of samples until ~300 samples; for training datasets with more than 300 samples, there is only a negligible decrease in the prediction error for rack inlet temperature. It should be noted that the prediction error for the models based on GPR framework is the smallest for a given number of samples in the training dataset; surprisingly, the prediction error is less than 1 °C even when there are only 50 samples in the training dataset. This could be due to the stochastic nature of GPR modeling, which may give it the capability to handle, and account for uncertainty in, small noisy

datasets. It therefore appears that the GPR framework is the best data driven model for smaller datasets.

4.2.1.4 Adaptability of Modeling Framework to Configuration Change

Given the rapid advances and changes in electronics and IT equipment, the computation and networking equipment in data centers have ever-shorter lifecycles and more rapid turnover. From the perspective of facility-side management, equipment refresh affects the power consumption of and heat distribution within data centers, and data centers must be designed so that they can accommodate these changes. From the modeling perspective, the challenge lies in developing a model that can adapt to successive configuration changes in data centers. Theoretically, such a model would require an infinite number of adjustable parameters (degrees of freedom), which is infeasible. Room-level models are typically specific to a given data center configuration, and would in most cases require recalibration of air flow patterns and thermal resistances (for lumped-capacitance models), constructing a new numerical model and updating boundary conditions (for CFD/HT models), or retraining the models with measurements from, or numerical simulations of, the new configuration in the training dataset (for data driven models).

In this section we study how data center re-configuration affects the prediction accuracy of the ANN, SVM and GPR data driven models developed here. Five scenarios, ranging from very slight to fairly drastic alterations to the data center room are considered (see Fig. 46).The first two scenarios have non-uniform load factor (LF) (see equation 3.4); $LF_{rowA} \rightarrow Non-uniform$ for case 1 and both LF_{rowA} & $LF_{rowB} \rightarrow Non-uniform$ for the second case. This is achieved by setting one and two racks, respectively, to a zero (0 kW)

load. Scenarios 3,4, and 5 involve physically removing a rack located at the end of an aisle, removing a rack located in middle of aisle, and removing two racks (one from aisle end and one from the middle), respectively. CFD/HT simulations were conducted to obtain results for all five modified configurations. The results from these simulations were compared with the predictions from the data driven models to evaluate their accuracy.

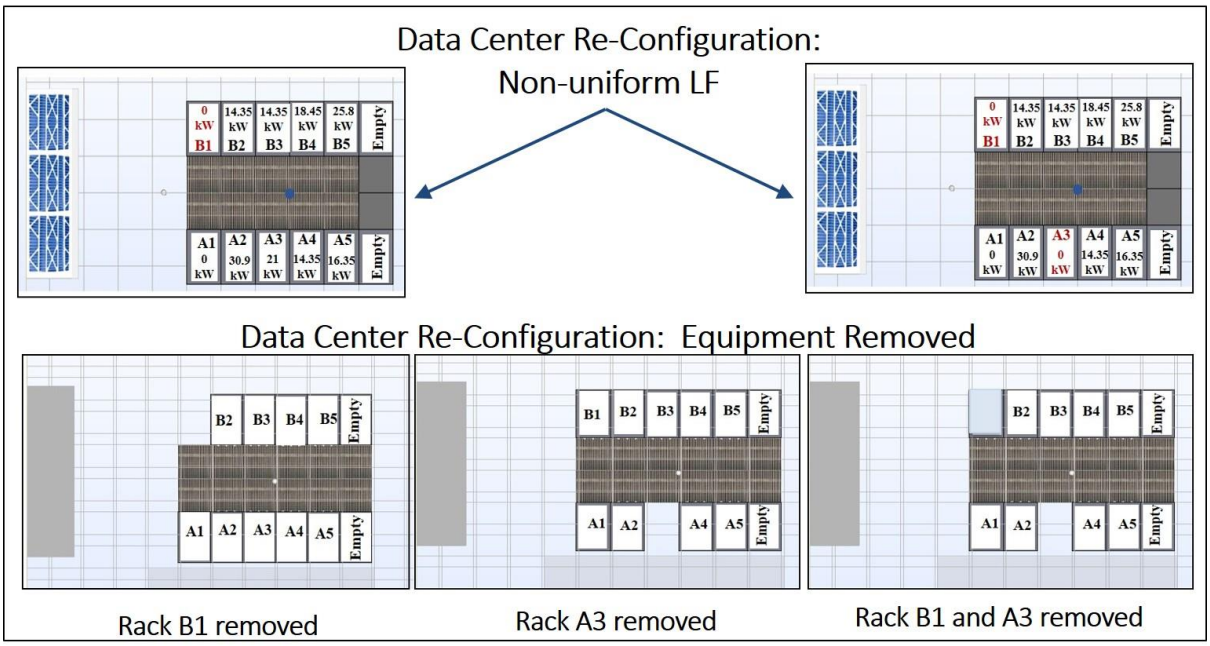
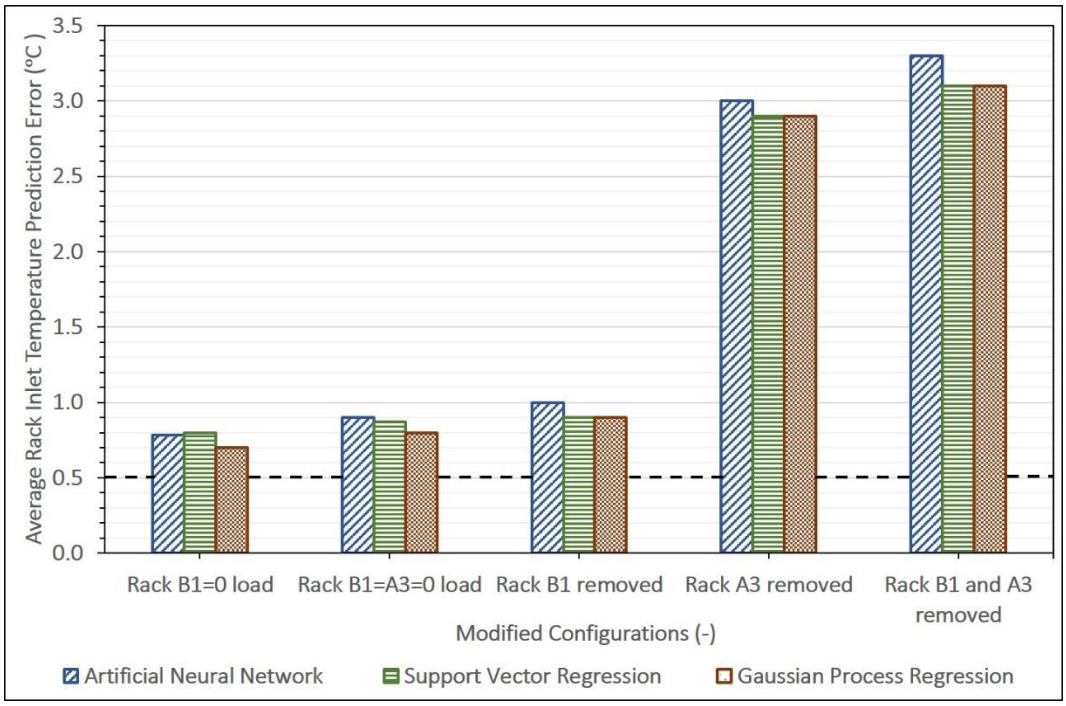


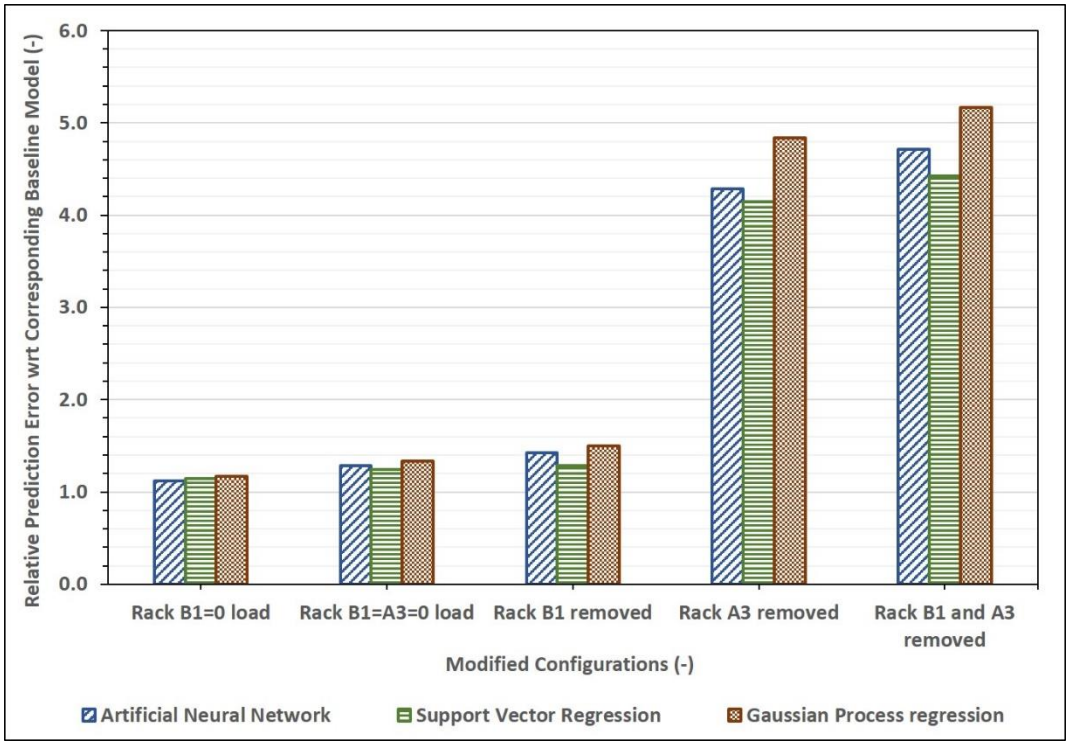
Figure 46 - Data Center Re-Configurations Tested

Figure 47(a) gives the overall average rack inlet temperature prediction discrepancy for ANN, SVR and GPR models for the five modified data center configurations. Figure 47(b) gives the error for the ANN, SVR and GPR models for these altered configurations normalized by the overall average errors of these models for the original data center configuration. The vertical axis thus represents the increase in prediction error when the data driven models are employed to make predictions for the altered scenarios; a value of unity (1) corresponds to no increase in the prediction error for the altered configuration

compared with the unaltered configuration, *i.e.*, the model can predict inlet rack temperatures for both configurations with the same accuracy. As seen in Fig. 47(a), the prediction error increases only slightly for non-uniform heat loads, remaining less than 1 °C in all cases. However, the prediction error is quite large for the scenarios (4 and 5) involving physically removing a rack from the middle of a row. This is because physically removing a rack from the middle of a row affects the airflow patterns on either side of that rack much more than just a redistribution of heat-load or removing a rack located at the end of an aisle. Among the three DDMs, the SVR model has the smallest increase in prediction error (*i.e.*, performance degradation) compared with the ANN and GPR models.



(a)



(b)

Figure 47 -(a) Prediction Error for ANN, SVR and GPR Models for Modified Data Center Configuration (b) Relative Prediction Error for ANN, SVR and GPR Models with respect to Corresponding Prediction Error (without Configuration Change)

4.2.2 Transient Modeling –Cooling Failure Scenario

4.2.2.1 Generating Training Data Set

The transient scenario considered here is a cooling failure scenario, where the chilled-water pump (CWP) fails while the CRAC blowers (CB) remain active, recirculating the air in the data center room (see Fig. 48). Here time 't' is the independent variable, while N_{CRAC} , $T_{a,ret}$ and $\dot{Q}_{IT\ room}$ are fixed parameters. The rack inlet temperatures at the 36 sensor points shown previously are monitored as the output variable in transient CFD simulation used for generating training data. As shown, the first 200 s of data (20 samples spaced 10 s apart) were used for training the models and testing the interpolative predictive capability of the model, while the next 300 s of data were used to test the extrapolative predictive capability.

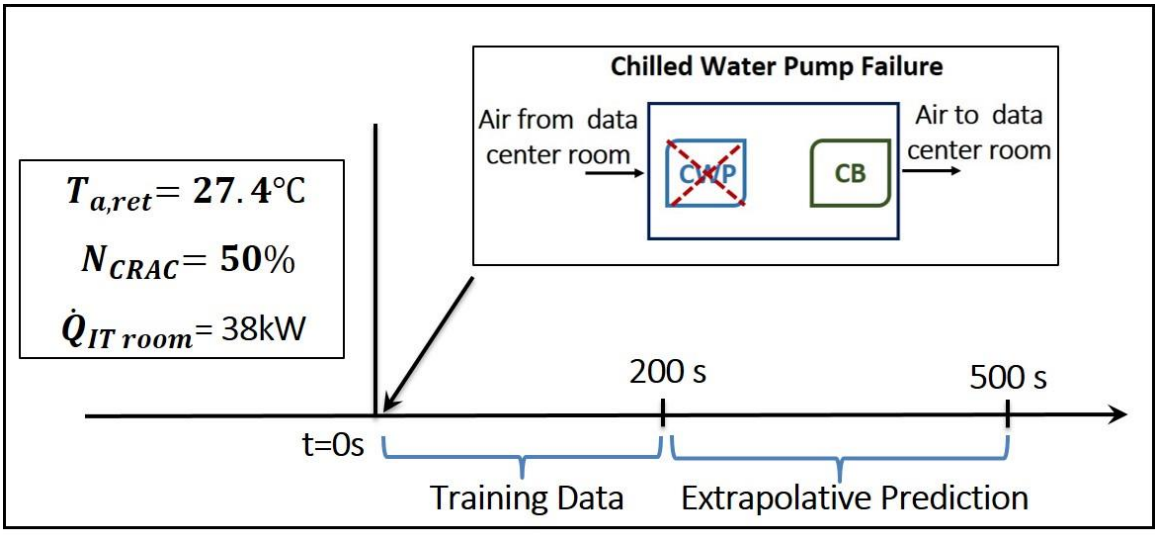


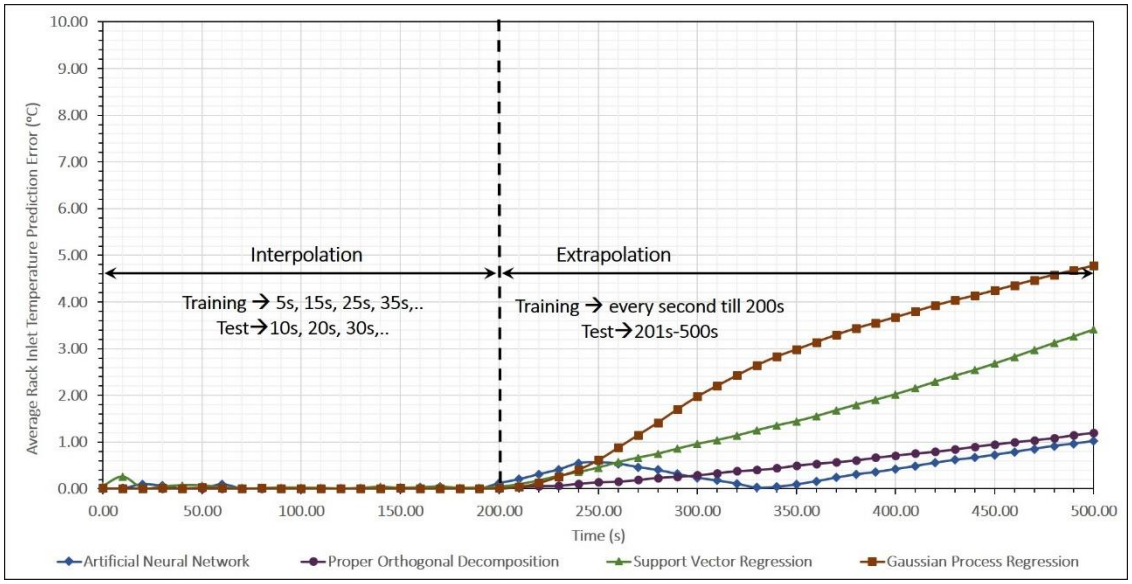
Figure 48 - Transient Scenario for Data Generation

4.2.2.2 Interpolative and Extrapolative Accuracy

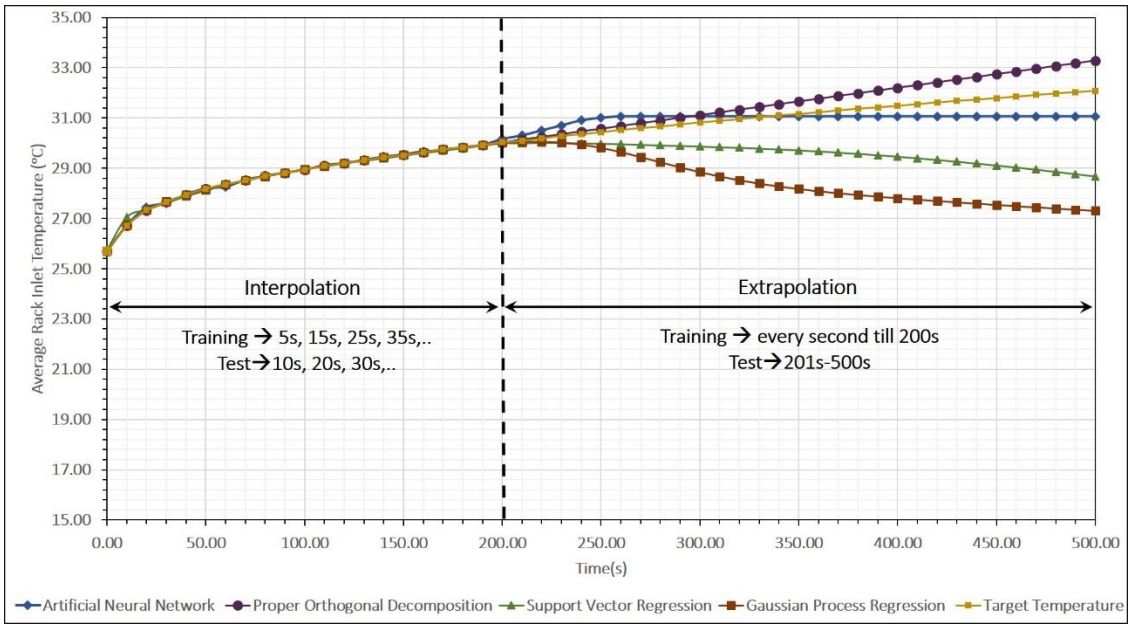
Figure 49(a) and 49(b) compare the absolute prediction error in and evolution of rack inlet temperature for the 500 s following cooling failure in the data center room for the ANN, POD, SVR and GPR models. As described earlier, the first 200 s of data were used for training the models and testing the interpolative accuracy, while the next 300 s of data were used to test the extrapolative accuracy. Every point on each of the curves represents overall error averaged over all 36 sensor locations. For testing the interpolative accuracy, the interrogation instances are separate from the ones used for training but still within the 200 s interval over which the models are trained, and are therefore interpolative in nature. The interpolative error associated with all the models is well below the measurement uncertainty of 0.5 °C. As can be seen in Fig. 49(b), the actual increase in the rack inlet air temperature is almost linear, and the models are able to capture this trend with very good accuracy.

The extrapolative accuracy for the data driven frameworks employed is compared in Fig. 49(a). The graph shows that all the models have temperature prediction errors that increase with time, starting from the end of the training data. The rate of increase is much higher for the predictions of the GPR and SVM models compared with those of the POD and ANN models. Interestingly, the temperature evolution trends in Fig. 49(b) show that only the POD model follows the monotonically increasing linear trend during the training period, while the other models show trends that make no physical sense. The ANN model has a sudden increase in temperature from 200 s to 250 s and thereafter maintains a constant temperature value, while SVR and GPR models actually have a monotonic decrease in temperature. For SVR and GPR, the choice of kernel function determines their asymptotic

behavior (here in extrapolation) of the developed models. In making predictions, these models look for input values in the training data set that were closest to the input values for the new interrogation points. For the two kernels used here, namely the radial basis and Gaussian squared exponential kernels for SVR and GPR models respectively, the kernel function value tends to zero ($e^{-\infty} \approx 0$; from equations 4.7 and 4.12) as the distance between the input values in the training set and that of the new interrogation point increases. This translates to the model tending to the mean value of the training data (here, ~28 °C) as the model progresses beyond the training data. If *a priori* knowledge of temperature evolution trend is available, better predictions may be possible by selecting a kernel function that grows linearly.



(a)



(b)

Figure 49(a) - Comparison of Interpolative and Extrapolative Prediction Error in Rack Inlet Temperature for ANN, POD, SVR and GPR Models (b) - Comparison of Transient Rack Inlet Temperature Prediction for ANN, POD, SVR and GPR Model

4.3 Summary

Two data driven models based on the SVR and GPR frameworks for predicting temperatures in data centers were developed in this work and results were compared to ANN and POD models developed in Chapter 3. The steady-state models were trained on the results from 300 experimentally validated CFD/HT simulations using the commercial CFD software Future Facilities 6Sigma Room. Verification tests indicated that the resultant three (ANN, SVR and GPR) steady-state models are reasonably accurate, with the GPR model having the lowest prediction error, 0.6 °C. For a given rack, the steady-state prediction error increases with height along the front face of rack; this may be due to the inability of the data driven models to capture recirculation effects.

The accuracy of these three modeling frameworks was also determined as a function of number of samples in the training dataset. The model based on the GPR framework had the lowest error for any given training dataset size, and was found to have surprisingly good accuracy, with a prediction error below 1 °C) for dataset containing only 50 samples.

Data driven models employing the ANN, POD, SVR and GPR framework were also developed for a transient scenario. These models were trained and tested using results from transient CFD/HT model simulating a cooling failure scenario. All four models have very low interpolative prediction errors, which are below the measurement uncertainty of 0.5 °C). The extrapolative errors are high, however, and appear to increase with temporal distance from the end of the training data. For the transient scenario under consideration, only the POD model can even predict the qualitative trend of the temperature, namely its

increase with time, after the cooling failure. It may be possible, however, to improve the extrapolation capabilities of SVR- and GPR-based models by selecting a kernel function if *a priori* qualitative knowledge of the trends in the evolution of the parameters are available

CHAPTER 5. COOLING ENERGY MODELING OF DATA CENTERS

5.1 Energy Usage for Data Center Cooling

As indicated in Chapter 1, in 2014 approximately 1% of total electricity consumption in the U.S. was for cooling data centers. In an air-cooled data center facility, the cooling infrastructure has three major elements: the chiller plant (employing a vapor compression refrigeration cycle), cooling towers and data center floor air-conditioning units.

Figure 50 shows the path for heat dissipation from a typical air-cooled data center to the surroundings. Cold air supplied by the CRAC blowers enters the data center room through perforated tiles, is drawn into the IT equipment by the server fans, then heated by the equipment, and returns to the CRAC unit. This warm air rejects heat to chilled water supplied by the building chiller via water-air heat exchangers in the CRAC unit. The condenser in the chiller unit ultimately rejects heat to the ambient via an air-cooled cooling tower. Variations of this cooling infrastructure design include replacing the water-cooled condenser and cooling towers with air-cooled condensers in the chiller vapor compression cycle that directly reject heat to the ambient, eliminating the chiller through use of direct expansion CRAC units. Recently, chiller-less data centers have been built which employ a combination of evaporative cooling and free cooling in regions where water and air are available at appropriate conditions.

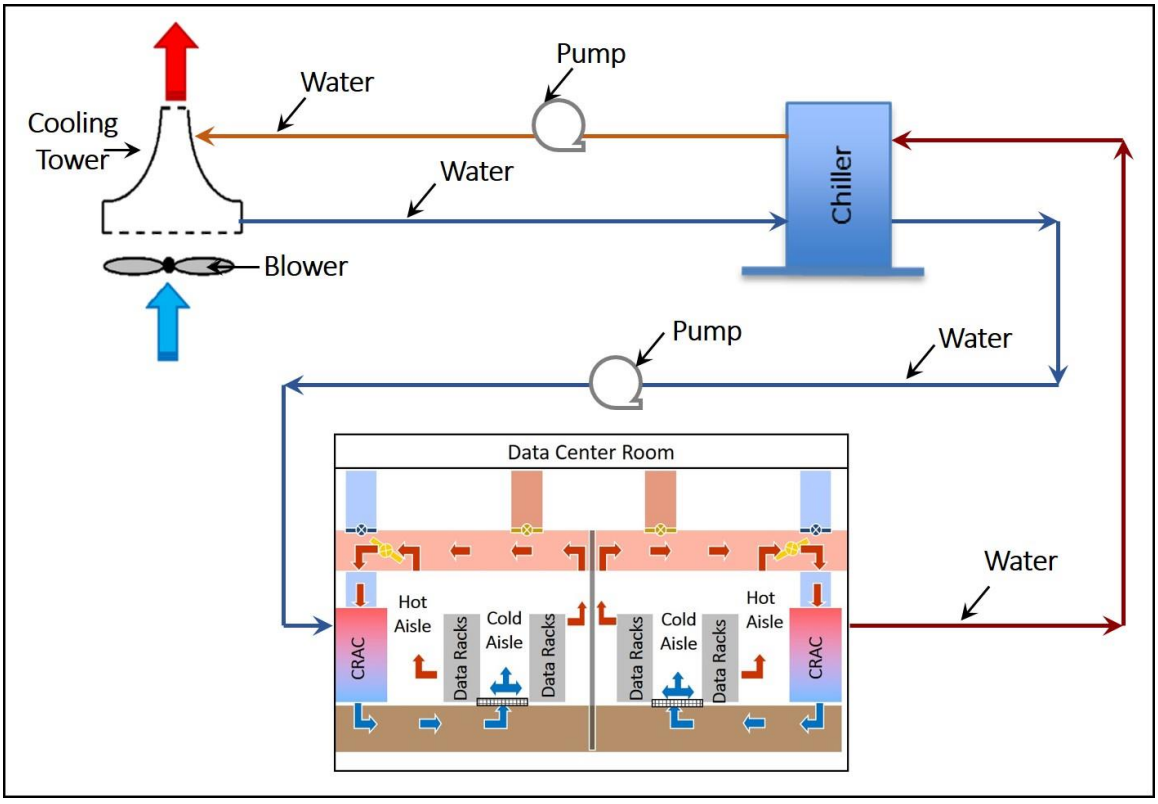


Figure 50 – Heat Flow in a Data Center

5.2 Cooling Energy Modeling

Optimization frameworks that consider cooling energy minimization in addition to thermal management constraints require modeling of the components of the cooling infrastructure and subsequent prediction of cooling energy consumption as indicated in Fig. 7 in Chapter 1. As such, many studies have focused on modeling components of the cooling infrastructure in data centers [6, 116-119]. These studies have provided accurate estimates of cooling energy consumption [6, 116], conducted parametric analyses to investigate the effect of varying component characteristics including chiller and CRAC on data center thermal and airflow profiles [117-119], and optimized the design of specific components [63]. The results from these studies have also informed decisions regarding the design and efficient operation of data centers. For cooling energy optimization, it is

essential that the cooling power be expressed as an explicit function of the control/manipulated variables, which are most commonly either CRAC blower speed (N_{CRAC}), or CRAC return temperature set-point $(T_{a,ret})$, or both.

Figure 51 shows the flow path for heat dissipated from within the data center at Georgia Tech to the ambient. The chiller unit being utilized is Trane Series R Rotary Chiller RTAA130GYT01A [120].

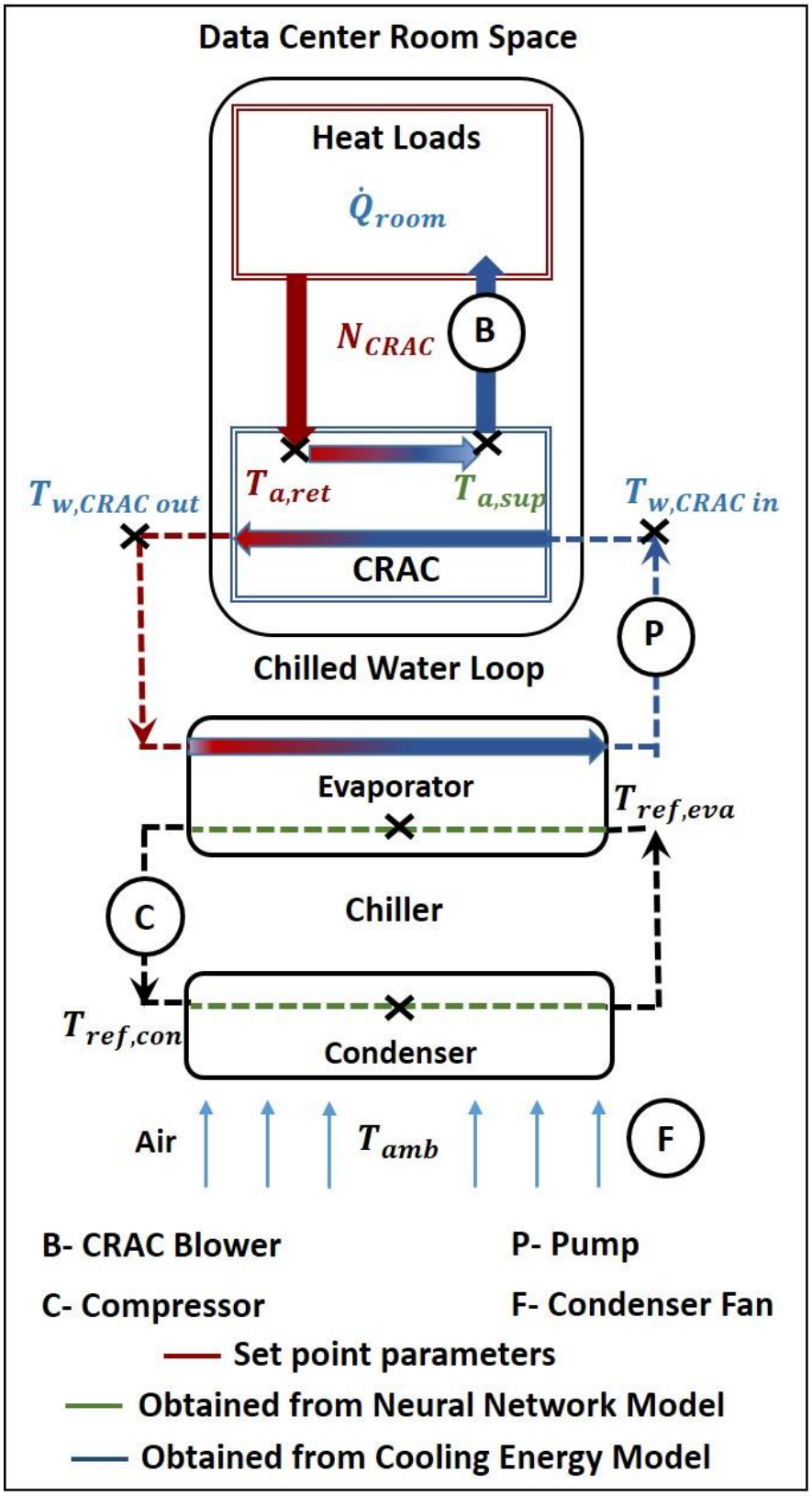


Figure 51 - Cooling Energy Modeling for Data Center Room

The portion of the energy consumed by data centers for cooling can be broadly divided into two parts: 1) energy consumed by chiller, which runs on a refrigeration cycle, and 2) energy consumed by fluid (air, water or refrigerant) transport components like CRAC blowers and server fans (equation 5.1).

$$P_{cooling} = P_{chiller} + P_{blower} + P_{server\ fans} \tag{5.1}$$

Note that the power from the liquid pumps, which is quite small, is not considered here. Also, energy consumed by the server fans $(P_{server\ fans})$ is a function of server fan speed and implicitly depends on rack inlet temperature via the fan's internal control logic. The contribution of server fans in cooling power consumption is accounted for in the energy modeling but the optimization framework does not directly target that segment.

Under steady/quasi-steady-state conditions, the following energy balances apply:

- Total heat load imposed on evaporator in the chiller:

$$\dot{Q}_{eva} = \dot{Q}_{IT\ room} + P_{blower} \tag{5.2}$$

- For the CRAC unit air flow:

$$\dot{Q}_{eva} = \dot{m}_a C_{p,a} (T_{a,ret} - T_{a,\mathrm{sup}}) \tag{5.3}$$

- For the chilled water loop:

$$\dot{m}_a C_{p,a}(T_{ret,a} - T_{\sup,a}) = \dot{m}_w C_{p,w}(T_{w,CRAC,out} - T_{w,CRAC,in}) \tag{5.4}$$

The water-air heat exchanger for the chilled water loop is a cross-flow heat exchanger and the effectivenesss is given by:

$$\varepsilon_{HEx,a-w} = 1 - \exp\left[\frac{1}{Cr} NTU^{0.22}(\exp[-(Cr)NTU^{0.78}] - 1)\right] \tag{5.5}$$

The *NTU* value is estimated based on manufacturers specifications for thermal conductance for the heat exchanger and heat capacities in use [6]; *Cr* is the capacity rate ratio. The effectiveness can be re-written as:

$$\varepsilon_{HEx,a-w} = \frac{T_{a,ret} - T_{a,\sup}}{T_{a,ret} - T_{w,CRAC,in}} = \frac{\dot{m}_w C_{p,w}(T_{w,CRAC,out} - T_{w,CRAC,in})}{\dot{m}_a C_{p,a}(T_{a,ret} - T_{w,CRAC,in})} \tag{5.6}$$

In equation 5.6, $T_{a,ret}$ is the temperature set-point, and $T_{a,\sup}$ is available from simulation results. Hence $T_{w,CRAC,in}$ can be determined if $\varepsilon_{HEx,a-w}$ for the heat exchanger is known. At design conditions, the chiller unit is set to maintain a 5.6 °C $(T_{w,CRAC,out} - T_{w,CRAC,in})$ temperature drop across the evaporator,. So all the required temperatures can be determined from the total heat load and heat exchanger specifications.

The chiller power consumption can be calculated using the manufacturer specified COP:

$$P_{chiller} = \frac{\dot{Q}_{eva}}{COP(T_{w,CRAC,in}, T_{amb})} = f(T_{a,ret}, N_{CRAC}) \tag{5.7}$$

The COP for a vapor-compression cycle is a function of evaporator and condenser temperatures, i.e., the chilled-water supply and ambient temperatures, respectively, for data center operation. Figure 52 shows how the chiller COP depends on the chilled water supply temperature and ambient temperature for the rotary chiller considered here [120].

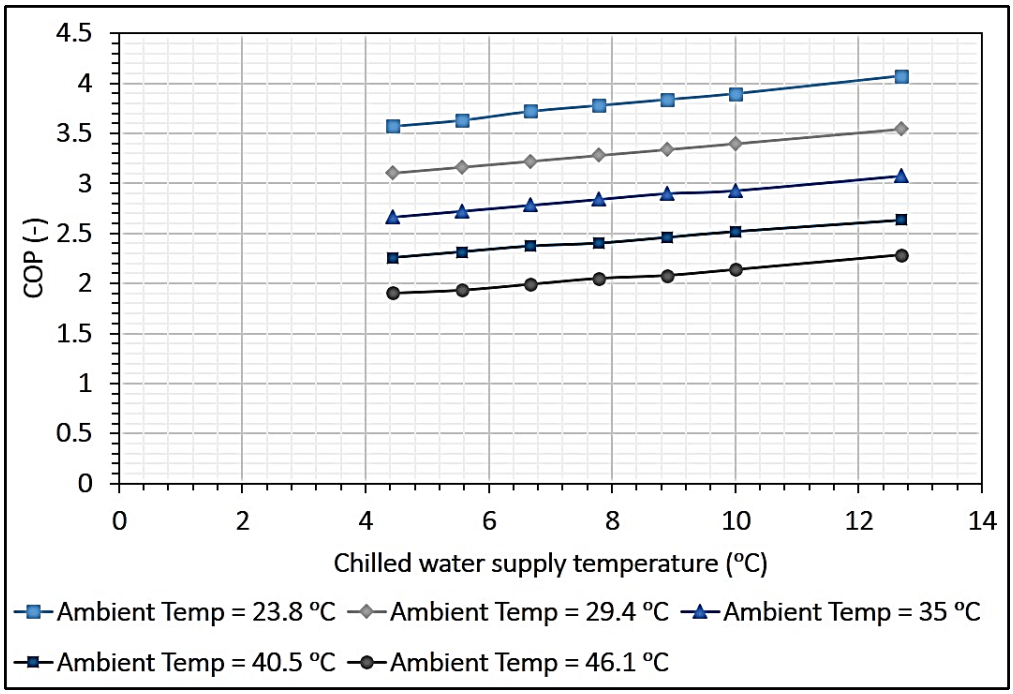


Figure 52 - Variation of COP with Chilled Water Supply Temperature and Ambient Temperature

Also, for a given chilled water-air heat exchanger in a CRAC unit, the higher the maximum allowed temperature for air supplied by the CRAC unit, the higher the maximum temperature for the chilled water supplied by chiller unit. Based on both of these trends, a higher CRAC supply temperature leads to a higher COP and therefore reduces the energy consumed by the chiller for a given data center room heat load. It has been estimated that

for every °F increase in chilled water temperature, the chiller COP improves by between 1% and 2.5%, where the exact magnitude of the improvement depends on the type of chiller. [121]

The second components of equation 5.1 (P_{blower}) is estimated from experimental measurements of the relationship between the blower/fan rotational speed and power consumed; the relation between the CRAC air flow rate $(\dot{m}_a)$ and blower speed is obtained from the blower curve supplied by the manufacturer.

CHAPTER 6. GENETIC ALGORITHM BASED COOLING ENERGY OPYIMIZATION OF DATA CENETRS

The primary function of the cooling infrastructure in data centers is to provide adequate cooling at all times and limit, if not eliminate, cooling failure-linked downtime. Concerns about reliability mean that the static, conservative set-points based on maximum IT capacity used at present lead to significant over-cooling of most data centers. Yet the energy consumed by data centers keeps on increasing, and is projected to increase to 75 billion kWh by 2020 [10]. As mentioned earlier, cooling infrastructure accounts for 30-50% of the total energy consumed by data centers [4]. These alarming trends and associated environmental impact are of great concern and have led to a new focus on *energy-efficient* thermal management of data centers.

This study presents the development and implementation of a Genetic Algorithm (GA)-based optimization framework, with the goal of minimizing the power required for data center cooling as a function of the optimization variables (cooling set-points), while ensuring that thermal management criteria (*i.e.*, rack inlet temperature limits) are satisfied. GAs, which are search algorithms based on evolutionary ideas of natural selection, use a population-based iterative approach in which multiple solution candidates participate and evolve a new population in each generation. Key components of this framework include computationally efficient and accurate models for thermal transport and cooling energy estimation. The Artificial Neural Network(ANN)-based model detailed in Chapter 3, [122], which is capable of predicting rack inlet temperature in nearly real time, is employed here.

The thermodynamics model of the data center cooling equipment developed in Chapter 5 is used to estimate cooling energy consumption.

The static optimization problem considers IT load distributions and cooling set-points in the data center room to simultaneously optimize two objectives: 1) minimize cooling power consumption, while 2) maximizing IT load. Three optimization scenarios, namely Room-Level IT Load, Row-Level IT Load Distribution and Rack-Level IT Load Distribution, employing IT load distributions at different spatial resolutions, are considered. Results for all three scenarios in the form of non-dominant solution series, *i.e.*, the Pareto-Front (PF), are presented. Experimental tests were conducted to determine the minimum cooling power required for these three scenarios when the overall load is varied from 40 kW to 160 kW; these results were compared to those from the GA-based optimization.

A quasi-static framework that aims to optimize cooling power consumption in the data center during operation is also developed. For a given IT workload distribution, the framework determines the most energy-efficient set-points for the cooling infrastructure, while the preventing the temperature from exceeding the recommended maximum. The framework was implemented for a test run of 7.5 h with a step change in room IT load every 30 min. The results are compared to two baseline operation cases where the CRAC set-points are kept constant with and without active CRAC return air temperature control. It should be noted that the data center room configuration is shown in Figs. 12(a) and 12(b) (in Chapter 2), and the location of rack inlet temperature sensors is given in Fig. 30 (Chapter 3).

6.1 Genetic Algorithm based Optimization

Genetic algorithms, which are a type of evolutionary optimization algorithm, are non-linear search and optimization techniques inspired by the biological processes of natural selection and survival of the fittest that are capable of handling single-, as well as multi-, objective optimization problems [123, 124]. GA-based optimization is a direct search procedure, and usually does not use gradient information in its search process. GAs have been used in a variety of heat transfer applications including optimization of absorption chiller designs [125], thermal comfort and energy consumption in buildings [126], and air distribution system design and operation [127]. An important aspect of GAs is that they can be used in conjunction with ANNs to solve the constrained multi-objective nonlinear optimization problems typical of data centers.

The proposed GA optimization framework applied to data center operation is summarized in Fig. 53. The methodology follows an iterative process starting with random generation of multiple candidate solutions, which form the initial population. The population size for each generation is determined based on the number of optimization variables for a given problem [128]. A larger population size enables a more thorough search of the solution space; it, however, also increases the computational time required. The fitness of each candidate solution is assessed based on objective function values (here, cooling power) and the feasibility is assessed using the constraint function values (here, rack inlet air temperature) associated with respective candidate solution (cooling set-point). The candidates for subsequent generations are selected using stochastic selection operators. In this study "tournament selection" was used as the operator, wherein two randomly chosen solutions from the evaluated population are compared and the "better" solution (in

terms of objective function values) is chosen to become part of the parent pool [128]. Candidates from the parent pool are included in the next generation either directly (inheritance), or via binary operations like cross-over and mutation. The process terminates when the prescribed criterion for the objective function(s) and constraint tolerance(s) are met, or there is no significant improvement between consecutive generations. Here, the optimum solution is a candidate (cooling set-points) with the minimum cooling power consumption possible without violating the constraints of rack inlet temperature for a given IT load distribution.

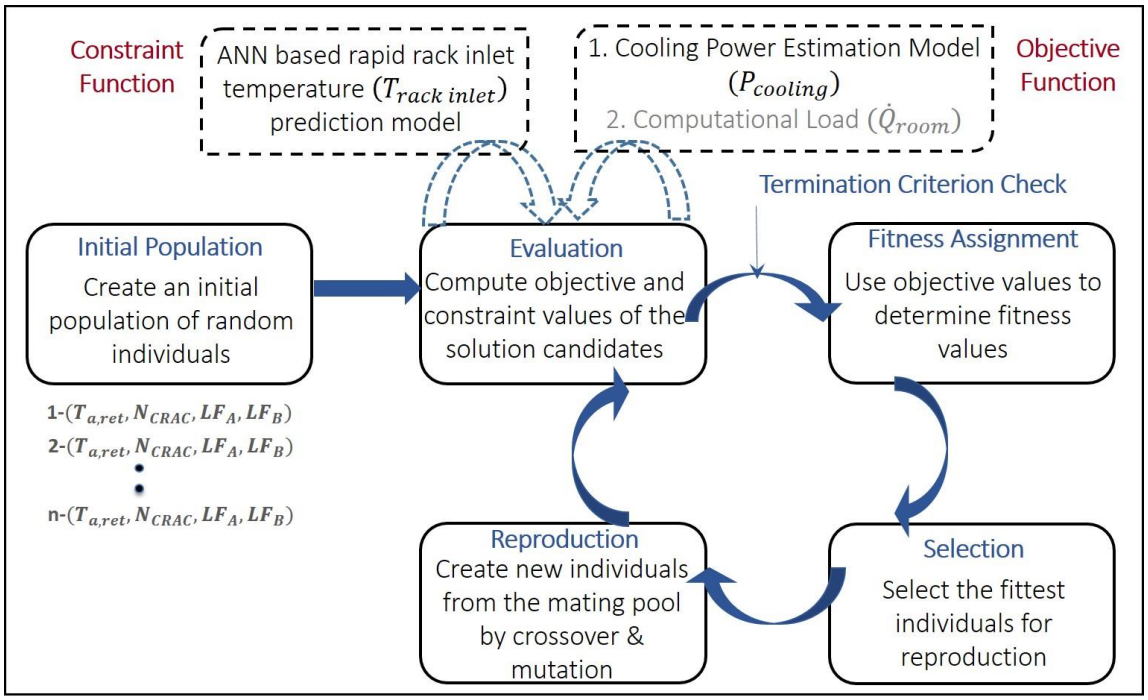


Figure 53 - GA Optimization Framework

6.1.1 Static Optimization

The two objective/cost functions for the static optimization in this study were minimizing cooling costs and maximizing the computational load for the given data center room. These two objectives conflict, which is typical for multi-objective optimization problems. The results of the static optimization are the ideal IT workload distribution among the racks and their corresponding cooling set-points to minimize cooling costs.

6.1.1.1 Optimization Problem Formulation – Static Optimization

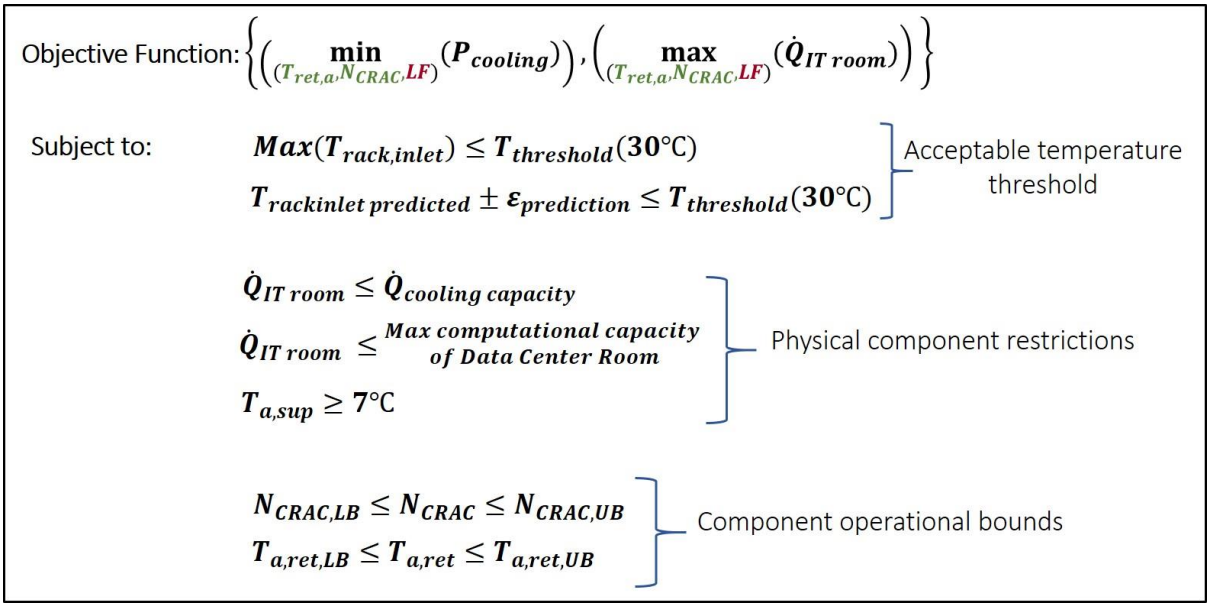


Figure 54 - Static Optimization Problem Definition

Fig. 54 shows the static optimization problem considered, wherein the optimization variables are the CRAC return air temperature $(T_{a,ret})$, CRAC blower speed (N_{CRAC}) and IT load distribution in the room represented by a Load Factor (LF). The Load Factor is that defined in Chapter 3 equation (3.4)

The thermal management constraints were formulated as optimization constraints and the rack inlet temperature predictions are obtained using the ANN-based model [122]. This means that the problem being formulated is an optimization with non-linear

constraints. Component restrictions and operational bounds for the CRAC unit have also been included in the formulation, as shown in Fig. 54. The maximum IT load for each rack in the data center room is indicated in Fig.12(b) (Chapter 2), and the total maximum room level IT load is ~164 kW. The optimization problem is formulated and solved using the commercial software package Matlab R2018.

Three different scenarios with varying levels of control of IT load distributions are considered: 1) Room-Level IT load; 2) Row-Level IT load distribution; and 3) Rack-Level IT load distribution. The set of optimization variables for each of these three scenarios is given in Table 13. Note that a separate rack inlet air temperature prediction model was implemented for each IT load distribution scenario.

Table 13- Static Optimization Scenarios

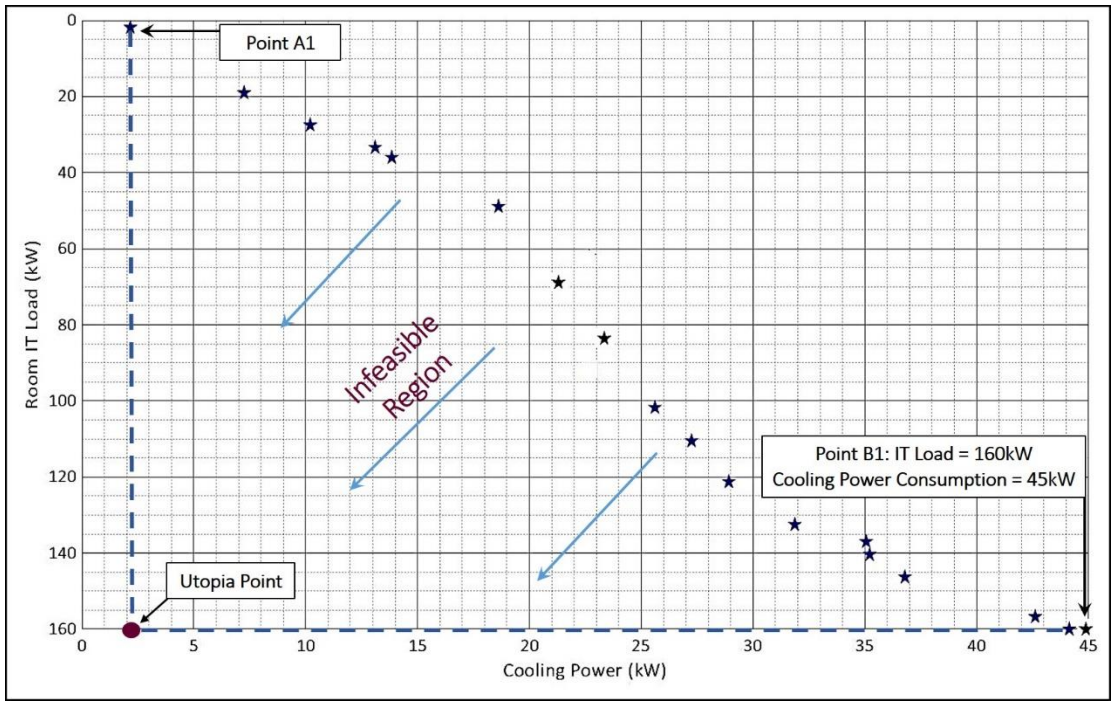
	Scenario	Explanation	Optimization Variables	Population Size for Optimization

1.	Room-Level IT Load	No granular control over IT load distribution and all the racks have uniform load factor	$\{N_{CRAC}, T_{a,ret}, LF_{room}\}$	100
2.	Row-Level IT Load Distribution	Row-wise control on IT load distribution and all racks in a given row (A or B) have uniform load factor	$\{N_{CRAC}, T_{a,ret}, LF_{rowA}, LF_{rowB}\}$	200
3.	Rack-Level IT Load Distribution	Rack-level control of IT load distribution with each rack having a distinct load factor	$\{N_{CRAC}, T_{a,ret}, LF_{A2}, LF_{A3}, LF_{A4}, LF_{A5}, LF_{B1}, LF_{B2}, LF_{B3}, LF_{B4}, LF_{B5}\}$	500

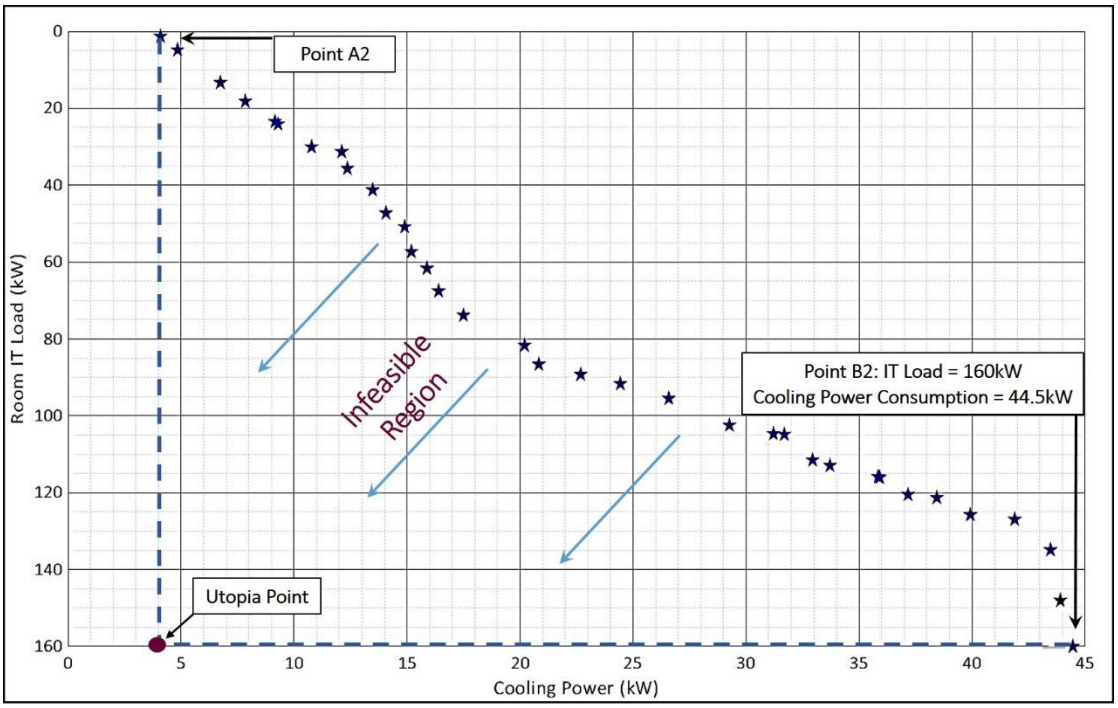
6.1.1.2 Results – Static Optimization

As mentioned previously, the static optimization problem represented in Fig. 54 is a multi-objective optimization problem resulting in a solution in the form of a Pareto Front (PF). A PF consists of a series of non-dominated candidate solutions, *i.e.*, no single candidate solution in the PF is superior to another solution with respect to all the objectives. Fig. 55 shows PF solution for the three scenarios: (1) Room-Level IT Load, (2) Row-Level IT Load Distribution, and (3) Rack-Level IT Load Distribution. As indicated in the Figure, solution candidates at A1, A2 and A3 consume the least cooling power, but for the lowest computational load in their respective PFs. By moving from points A1, A2, and A3 to B1, B2 and B3, respectively along the PF, each solution candidate can accommodate a higher room IT load, up to the maximum of 160 kW, at the "cost" of higher cooling power. The infeasible region below and to the left the PF represents the solution candidates that violate the rack inlet air temperature constraint. For all three scenarios, the region above and to the

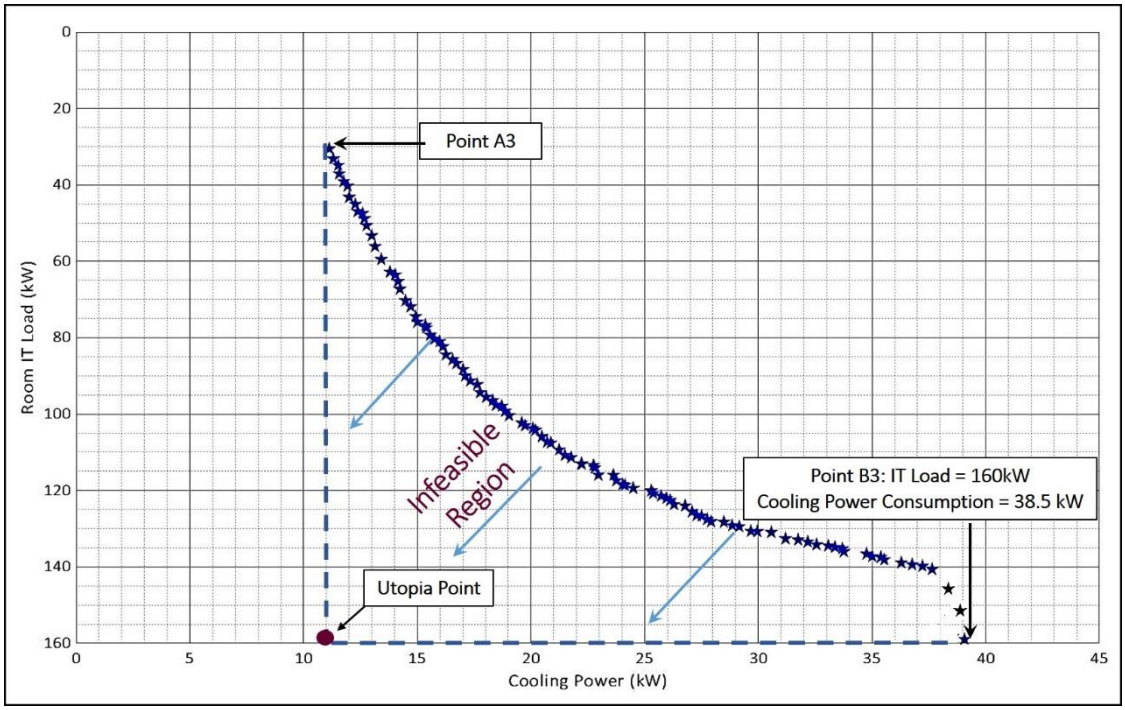
right of PF is comprised of sub-optimal solutions, *i.e.* solution candidates that satisfy all the constraints but consume more cooling power than the minimum required for a given load distribution.



(a)



(b)



(c)

Figure 55 - (a) Pareto Front for Static Optimization with Room Level IT Load (b) - Pareto Front for Static Optimization with Row Level IT Load Distribution (c) - Pareto Front for Static Optimization with Rack Level IT Load Distribution

Table 14 shows the cooling costs associated with these three load distribution scenarios. The cooling power consumption is lowest for scenario 3 with rack-level IT load distribution, and highest for scenario 1 with room-level IT load for the same overall room-level IT load of 160 kW.

Table 14- Cooling Cost for Static Optimization Scenarios; Room IT Load =160kW

Scenario	Cooling Cost (kW)	Total Room IT Load (kW)
Room-Level IT Load	45	160
Row-Level IT Load Distribution	44.5	160
Rack-Level IT Load Distribution	39	160

The optimization set-points determined using GA for the case when total room IT load is 160 kW and the rack-level IT load distribution is employed were implemented in the data center room to obtain the corresponding steady-state rack inlet air temperatures at all sensor locations (Fig. 56). The parameters implemented are - $\left\{N_{CRAC}, T_{a,ret}, LF_{A2}, LF_{A3}, LF_{A4}, LF_{A5}, LF_{B1}, LF_{B2}, LF_{B3}, LF_{B4}, LF_{B5}\right\} \rightarrow \{90, 26.6, 1, 1, 1, 0.9, 1, 1, 1, 1, 1\}$. The results clearly demonstrate that this scenario does not violate the temperature thresholds.

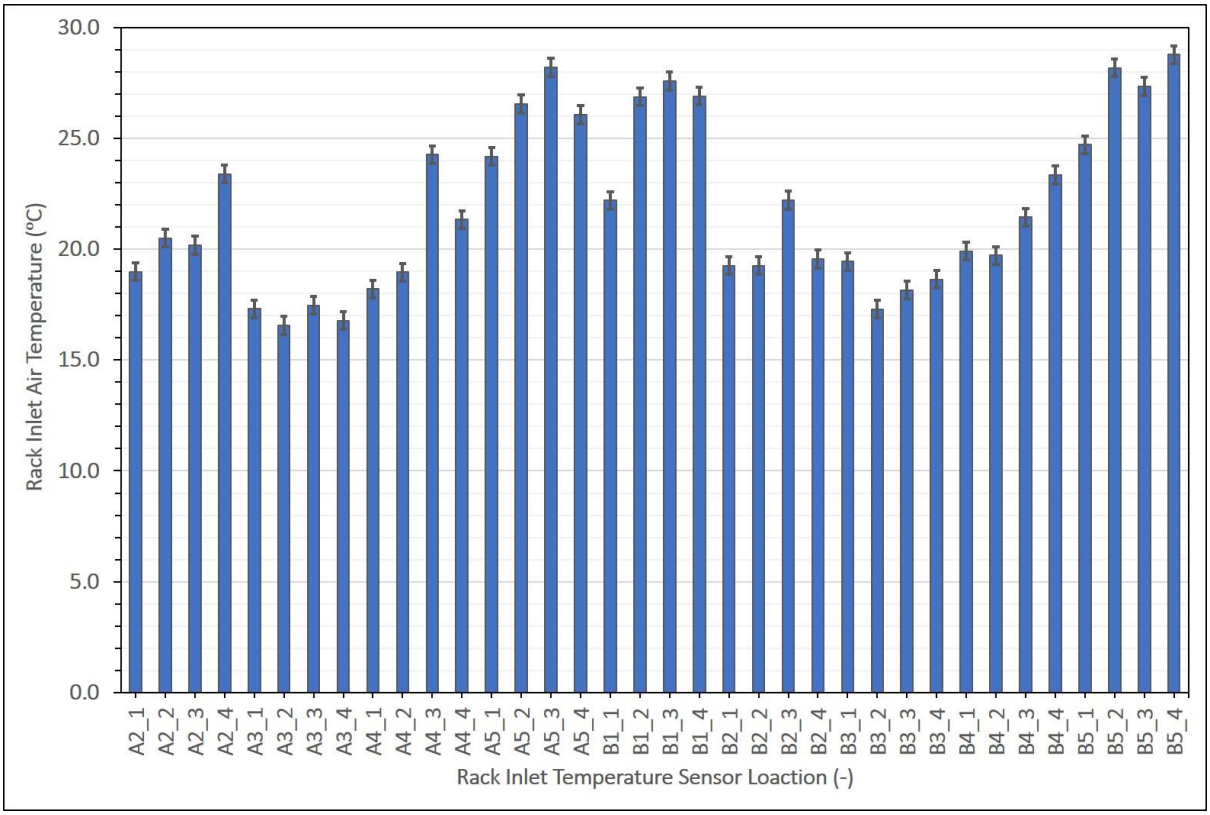


Figure 56 - Rack Inlet Air Temperature Distribution for Static Optimization based on Rack

Multiple optimization runs were conducted where the room-level maximum required computational capacity was varied from 40 kW to 160 kW in steps of 20 kW for the three IT load distribution scenarios. Fig. 57 shows the cooling cost as a function of room-level IT load for all three load distribution scenarios. It can be seen that for any given room-level IT load, optimization involving rack-level IT load distribution gives the minimum cooling power consumption. Moreover, the savings in cooling power consumption are especially significant when the IT load in the room is much lower than the full load capacity. Hence, controlling the distribution of IT loads at the individual rack level, namely scenario 3, provides greater flexibility and precision in accommodating the required computational load (without violating temperature constraints) compared to the other two IT load distribution scenarios. This level of granularity makes it possible to

distribute computational loads while accounting for spatial variations in temperature and airflow within the data center.

It should be noted that optimization with rack-level IT load distribution involves a larger number of optimization variables and thus larger population sizes and longer computational times (see Table 15). However, the increase in computational time is non-critical for static optimization because this type of optimization is *static*. At the highest load of 160 kW, all the racks are running at full capacity (LF = 1), and as such static optimizations of any of the three IT load distribution scenarios yield almost the same values for cooling set-points.

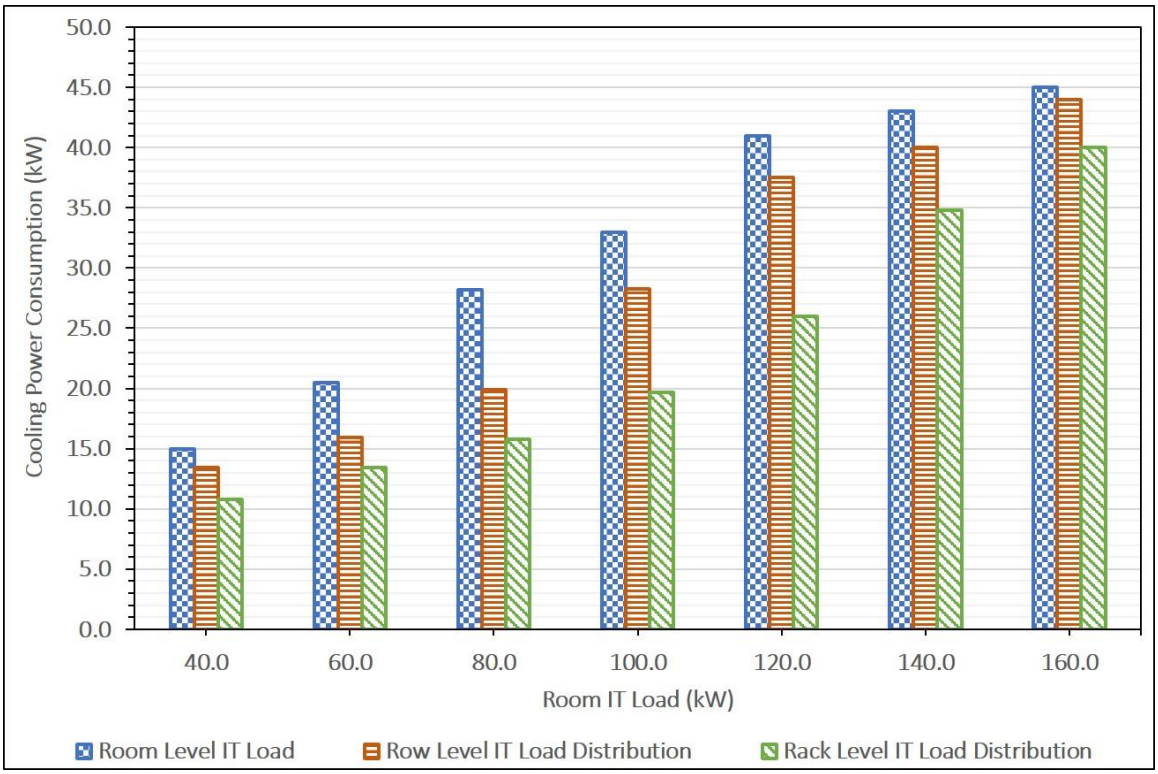


Figure 57 – Cooling Power Consumption as a Function of Room IT Load for Different IT-Load Distribution Schemes

Table 15 – Computational Time for Static Optimization Scenarios

Scenario	Average Computational Time Required (s)
Room-Level IT Load	12.3
Row-Level IT Load Distribution	23.2
Rack-Level IT Load Distribution	60.2
-The computational time reported for each scenario is an average over five optimization runs -All optimization is run on a machine with 32 Gb RAM and Intel© Xeon© CPU 2.60GHz	

6.1.2 Quasi-static Optimization

In reality, the IT workload in data centers varies over time. Static set-points for the cooling infrastructure based on the maximum room IT load capacity therefore lead to conservative operation and over-cooling. A survey in 2013 indicated that 90% of data centers operate at temperatures under the set-point of 24 °C, suggesting that they are overcooled and wasting energy [4]. Implementation of quasi-static actuation of cooling infrastructure based on results obtained from GA based optimization is described next.

6.1.2.1 Optimization Problem Formulation – Quasi-Static Optimization

Fig. 58 shows the quasi-static optimization case considered where $T_{a,ret}$ and N_{CRAC} are the optimization variables. The objective of optimization framework in this study is to minimize the cooling power consumption of the data center room while the rack inlet temperature limits are the constraints. As in the case of static optimization, the ANN-based model is used to predict rack inlet air temperatures. The component restrictions and operational bounds for the CRAC unit are also been included in this formulation as shown in Fig. 58.

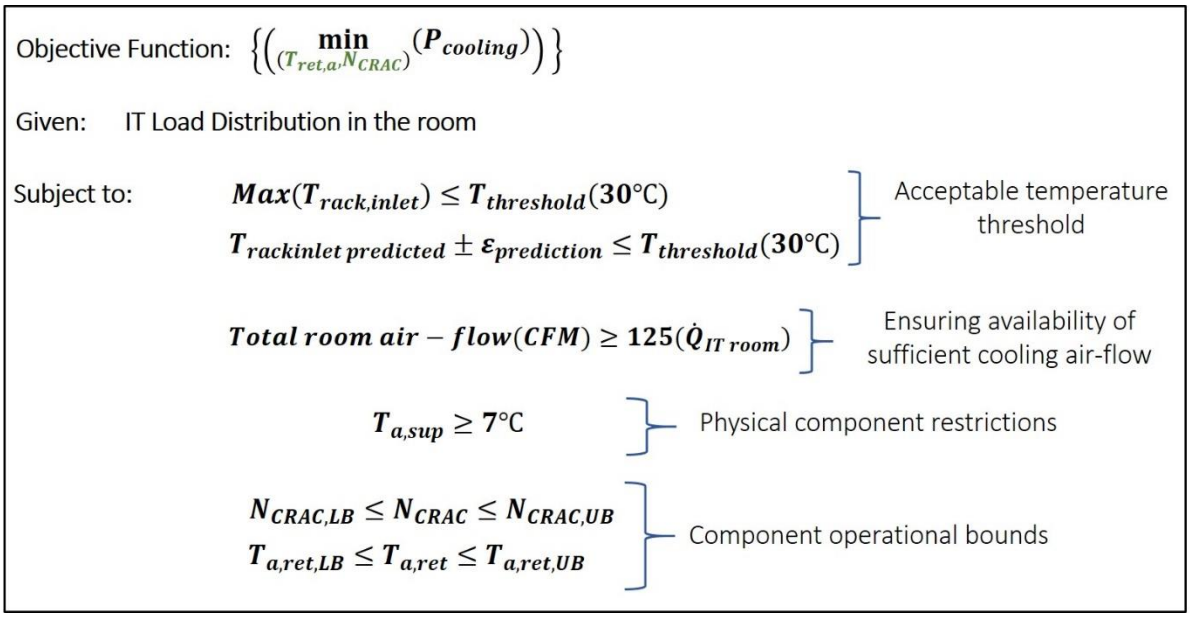


Figure 58 – Quasi-Static Optimization Problem Formulation

The quasi-static optimization framework was implemented in the data center for a duration of 7.5 h (= 450 min) where a step change in the IT load was applied every 30 min. Latin Hypercube Sampling (LHS), a statistical method for generating a near-random sample of parameter values, was employed to generate the room IT load profile depicted in Fig. 59. It should be noted that row-level IT load distribution was considered while generating the IT load profile so that all the racks in a given row have the same LF_{row}. The length of the control cycle was determined by conducting tests to determine "thermal time constant" for the Data Center Laboratory, *i.e.*, the time required for rack inlet temperatures to reach steady-state after changing the cooling set-points and IT loads.

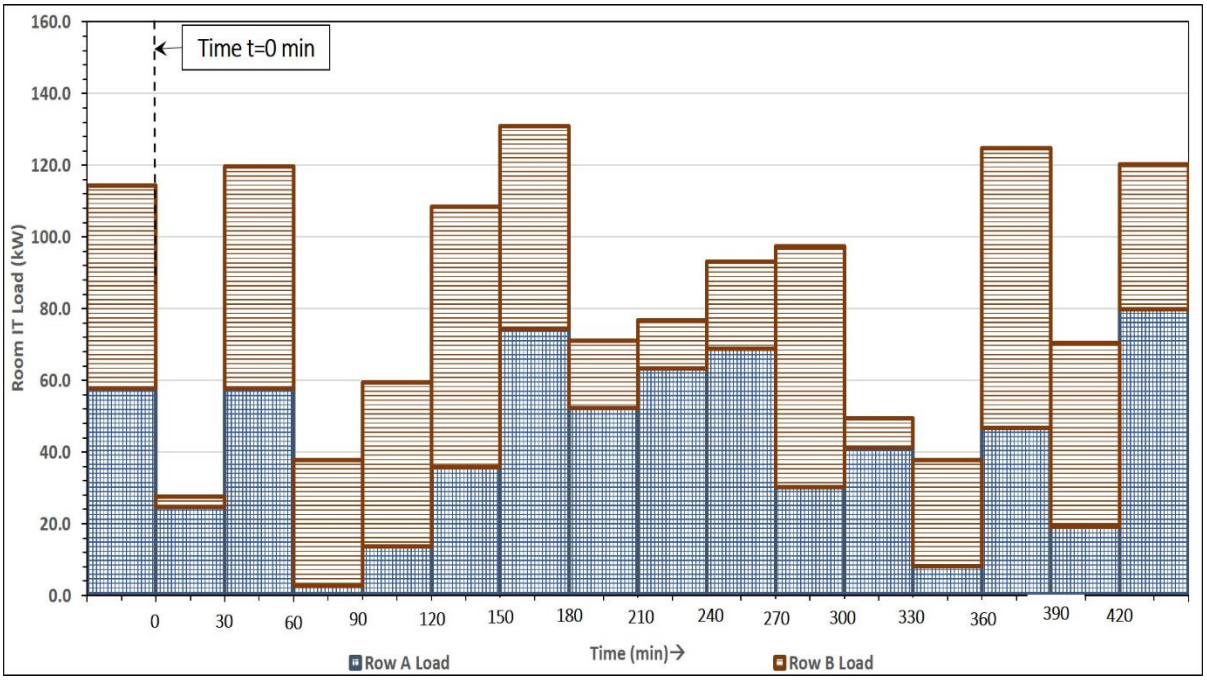


Figure 59 - Room IT Load Profile for Quasi-Static Optimization Test Case

The CRAC unit employed in this study operates on return-air temperature control.

Three different cases based on different methods for controlling the cooling set-points were considered: 1) GA-Based Optimization of Cooling Set-Points; 2) Constant Cooling Set-Point with CRAC Return Air Temperature Control; and 3) Constant Cooling Set-Points with No CRAC Return Air Temperature Control. For cases 2 and 3 the cooling set-points $(N_{CRAC}, T_{a,ret})$ were determined on the basis of the maximum IT load capacity of the room (= 165 kW) for the room, and kept constant through-out the experiment. The CRAC blower speed (N_{CRAC}) was determined so that ASHRAE (American Society of Heating Refrigeration and Air-Conditioning Engineers) specifications [2] for supplying 125 CFM/kW of cooling air were satisfied, while the return air temperature set-point $(T_{a,ret})$ was selected to satisfy rack inlet air temperature constraints. For case 2, the built-in return air temperature control for the CRAC unit was active, and controlled the CRAC

supply air temperature based on the room IT load and cooling airflow supplied. Case 3 represents the scenario with no temperature control of the CRAC unit where both the CRAC return and supply temperature are held constant; the CRAC supply air temperature was determined by a thermal balance of the data center room. For the GA-based quasi-static optimization framework developed here, the cooling set-points were determined and implemented at the beginning of every control cycle (*i.e.*, every 30 min). Table 16 summarizes these three cases and gives the constant cooling set-points for cases 2 and 3.

Table 16 – Test Cases for Quasi-Static Optimization

No.	**Case**	**Cooling Set-Points** $(N_{CRAC}, T_{ret,a})$	**CRAC Supply Temperature** $(T_{\sup,a})$
1.	GA-Based Optimization of Cooling Set-Points	-Different for each control cycle	-Varying to maintain the return air temperature set-point
2.	Constant Cooling Set-Point with CRAC Return Air Temperature Control	-Constant for entire test duration $(N_{CRAC} = 100\%; T_{ret,a} = 22°C)$	-Varying to maintain the return air temperature set-point
3.	Constant Cooling Set-Points with No CRAC Return Air Temperature Control	Constant for entire test duration $(N_{CRAC} = 100\%; T_{ret,a} = 22°C)$	Constant for entire test duration $(T_{\sup,a} = 12°C)$

The test run implementing the load profile in Fig. 59 was conducted for two cases, namely GA-based optimization of cooling set-points and Constant Cooling Set-points with CRAC return air temperature control. The third case was not experimentally implemented due to constraints posed by the control system of the actual CRAC unit; the associated cooling cost was estimated, however, using model described in Chapter 5.

6.1.2.2 Results – Quasi-Static Optimization

The rack inlet temperatures for all nine racks (four temperature points for each rack: *cf.* Fig. 30, Chapter 3) were monitored throughout the experiment for both cases. Figure 60 plots the maximum rack inlet temperature as a function of time. It can be seen that the maximum rack inlet temperature remains at or under the prescribed rack inlet temperature threshold of 30 °C for both cases, *i.e.*, when the cooling set-points are determined using developed GA-based optimization, and the case with constant cooling set-points and CRAC return air temperature control. The maximum rack inlet temperature when using constant cooling set-points is lower than that with GA-based cooling set points at all times because the set-points are more conservative in the former case.

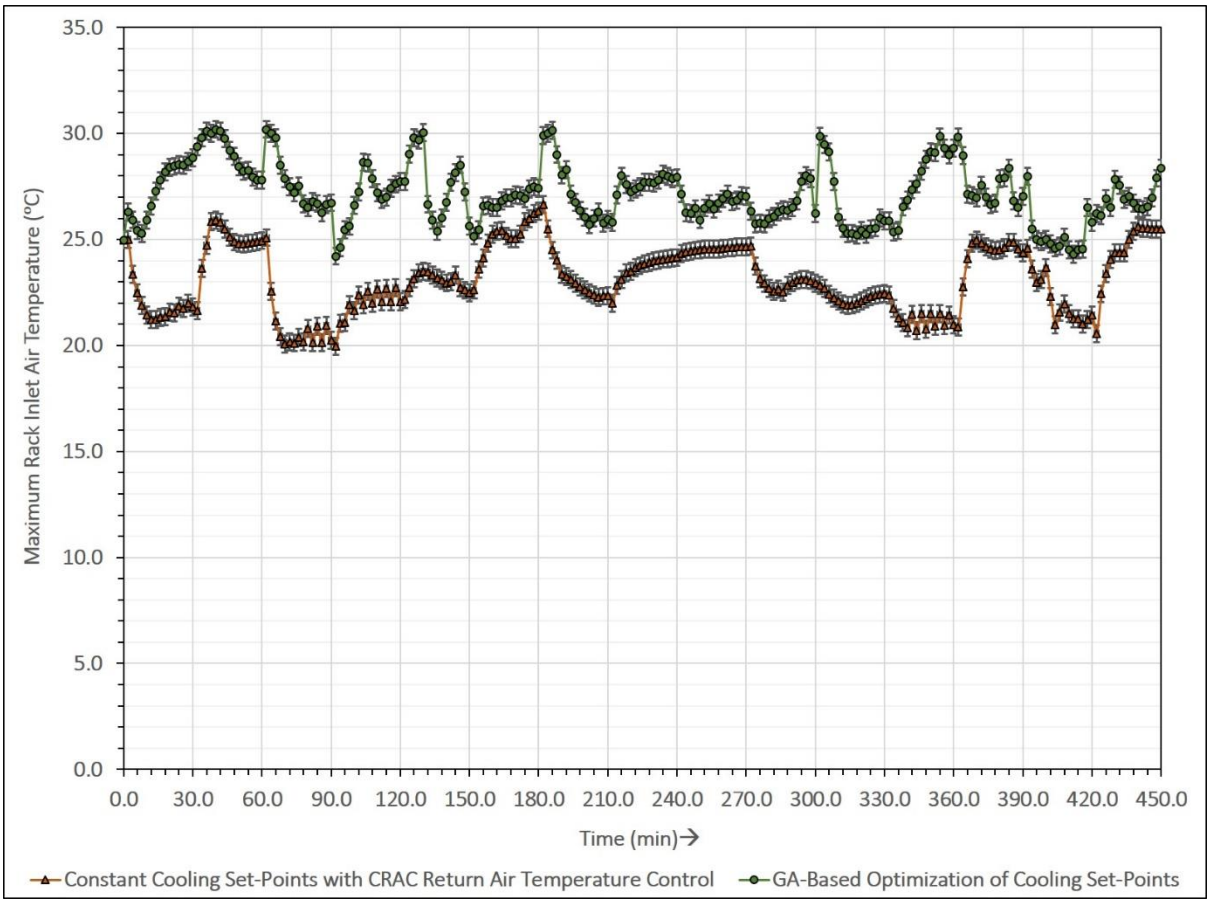


Figure 60 - Maximum Rack Inlet Temperature as a Function of Time for Case 1 (GA-Based Optimization) and Case 2 (Constant Cooling Set-Points with CRAC Return Air Temperature Control)

The cooling power consumption for each of the three cases, estimated using the procedure detailed in Chapter 5, is presented in Fig. 61. The cooling power consumption is lowest with GA-based optimization for each control cycle. The two factors contributing to this reduction in power consumption for the GA-based optimization framework (*vs.* using constant set-points with CRAC return control) are a lower CRAC blower speed (and therefore lower CRAC blower power consumption), and a higher return air temperature set-point (and therefore higher COP for chiller vapor compression cycle). It can be seen in Fig. 61 that the difference in cooling power consumption is not very high when the heat load in the room in very high. It was observed that the cooling set-points determined by the GA-based optimization depended on the two row-level load factors, namely $\boldsymbol{LF_A}$ and $\boldsymbol{LF_B}$, in addition to the overall room IT load. Specifically, for the same IT load, a higher load factor for row B, $\boldsymbol{LF_B}$, resulted in more conservative set-points to satisfy the temperature constraints. So the GA-based optimization gives higher energy savings compared with case 2 when the load is distributed such that $\boldsymbol{LF_A > LF_B}$. The cooling power consumption for the case with constant cooling set-points and no return air control (case 3) is constant over the entire test run, as shown by the dashed line. It should be noted that case 3 essentially represents a situation with a constant CRAC supply temperature of 11 °C, a very conservative set-point. Table 17 compares the cooling energy consumption for all three cases for the test run lasting 7.5 h. For this test run, the cooling energy consumed by the GA-based optimization framework was half that of case with cooling set-

points with no return temperature control (case 3) and 20% lower than the case with constant cooling set-points and return temperature control (case 2).

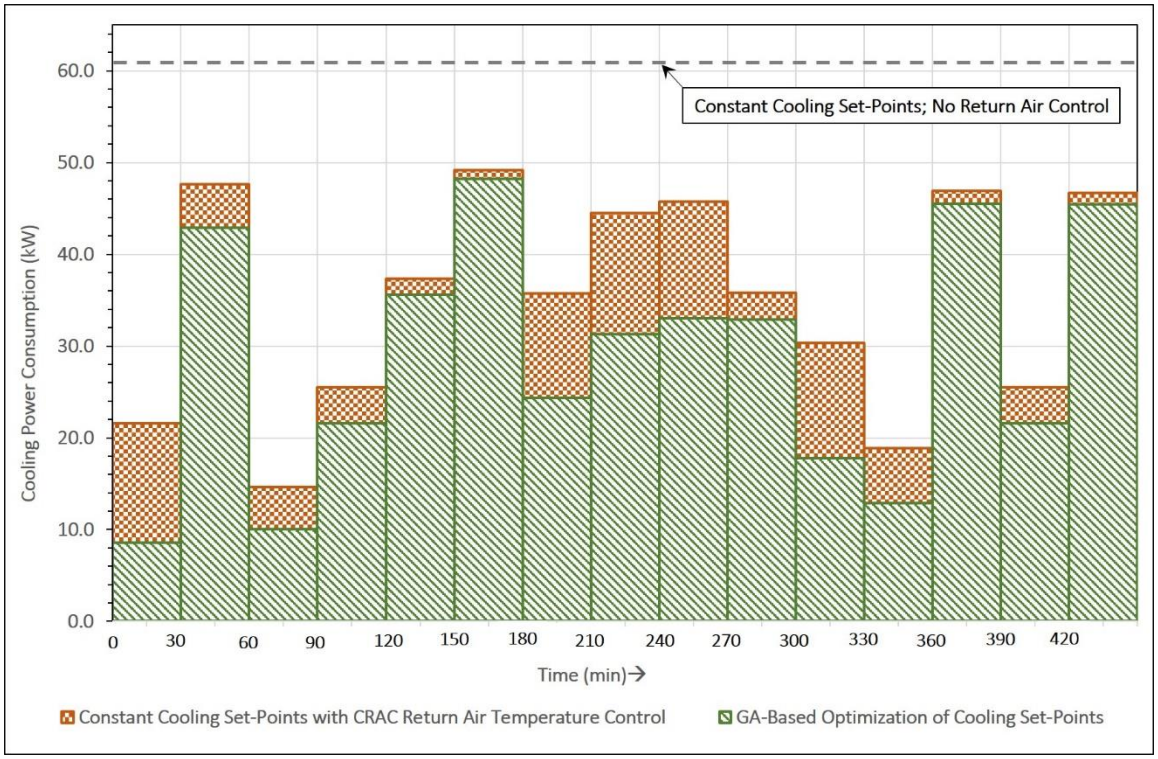


Figure 61 - Cooling Power Consumption as a Function of Time for Case 1 (GA-Based Optimization), Case 2 (Constant Cooling Set-Points with CRAC Return Air Temperature Control) and Case 3 (Constant Cooling Set-Points with No CRAC Return Air Temperature Control

Table 17 – Comparison of Cooling Energy Consumption for Test of 7.5 hours

No.	Case	Cooling Energy Consumption for test run of 7.5 hours (kWh)
1.	GA-Based Optimization of Cooling Set-Points	216
2.	Constant Cooling Set-Point with CRAC Return Air Temperature Control	263
3.	Constant Cooling Set-Points with No CRAC Return Air Temperature Control	457.5

6.2 Summary

This work developed both static and quasi-static GA-based frameworks for optimizing the cooling energy consumed by a data center. Rapid predictions of rack inlet

temperatures and cooling power consumption, which are key components of the optimization framework, are obtained by employing an ANN-based data driven and thermodynamics-based models, respectively.

The multi-objective, static GA-based optimization framework was developed to determine optimum IT load distributions and cooling set-points to minimize cooling power while maximizing the IT load in the room. Three levels of IT load distributions were considered: Room-Level IT Load, Row-Level IT Load Distribution and Rack-Level IT Load Distribution. The results show that Rack-Level IT Load Distribution, which offers the greatest flexibility and control, had the lowest cooling power consumption for any value of total IT load in the data center room.

The objective of the quasi-static optimization framework was to minimize cooling power consumption for data center operation while staying within temperature limits. The framework was implemented for a test run of 7.5 h where the IT load was varied in a pre-determined manner every 30 min and new cooling set-points determined by GA-based optimization were applied. The results were compared to cases where the cooling set-points, determined based on the maximum IT load capacity of the room, are held constant either with or without CRAC return temperature control. The GA-based optimization approach reduces cooling energy consumption by 20% compared with the case of constant cooling set-point and active return air temperature control, and by 52% compared with the case of constant cooling set-point with no return air temperature control.

CHAPTER 7. CONCLUDING REMARKS

As stated at the outset, the goal of this Ph.D. thesis was the development of an overall framework for cooling energy optimization in data centers (see Fig. 62). Different chapters of this dissertation have detailed development and implementation of key components of this overall framework.

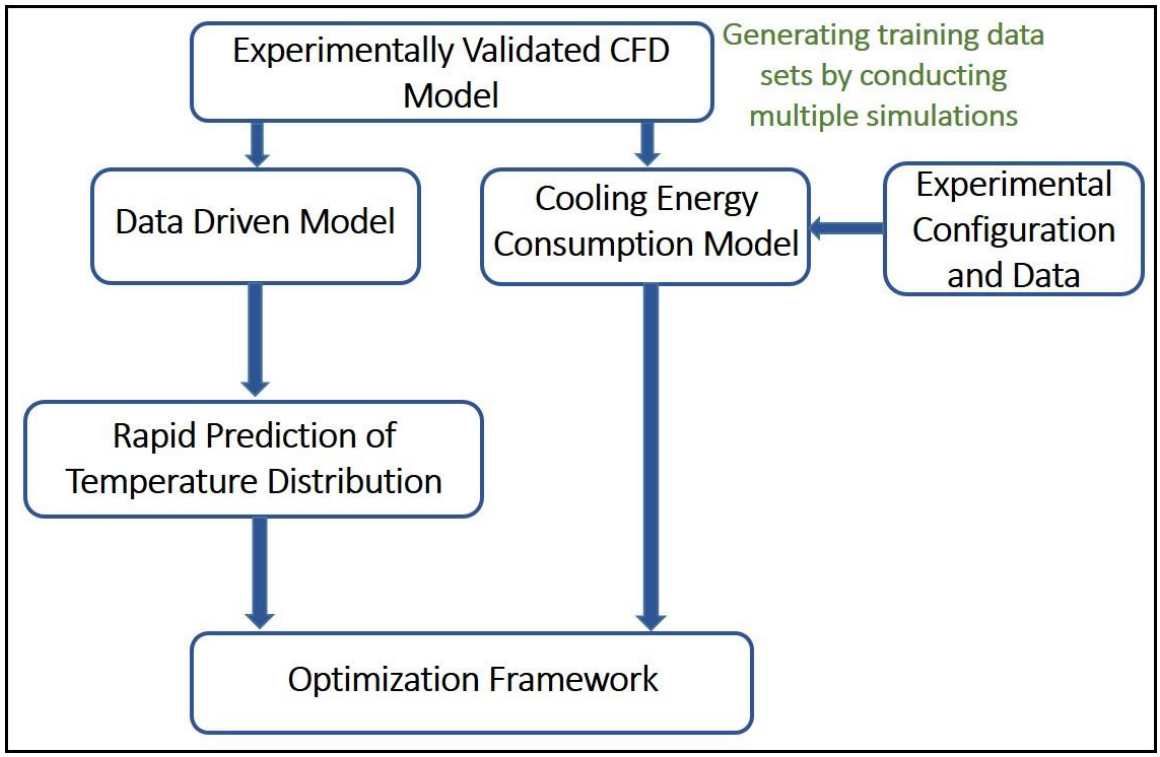


Figure 62 - Framework for Data Center Cooling Energy Optimization

Chapter 2 presents development of an experimentally validated room-level CFD model for the Data Center Laboratory at Georgia Tech. The model was developed using the finite-volume software package Future Facilities 6SigmaDCX, employing a pressure-based solver. The average overall discrepancy between the numerical predictions and experimental measurements was found to be less than 4% for total tile flow rate and less than 1.7 °C for rack inlet temperature. Parametric simulations conducted using this physics-

based numerical models were further used to generate training datasets for data driven modeling frameworks explored in this thesis.

Chapters 3 and 4 explore and compare the performance of four data driven modeling approaches capable of running in near real-time and therefore can be incorporated in the overall optimization framework. DDMs are particularly suitable for modeling data centers, given the non-linear nature of thermal transport, complexity of system operations and large number of associated metrics. Each of the four methods, namely Artificial Neural Networks (ANN), Support Vector Regression (SVR), Gaussian Process Regression (GPR) and Proper Orthogonal Decomposition (POD), possess distinctive features and attributes and were thus selected to gain a better understanding of the relative merits of each method as applied to data center thermal modelling.

The data used for training and analysis were obtained by performing 300 offline numerical simulations using CFD model developed in Chapter 2. For steady-state modeling, the multidimensional input parameter space for the simulations consisted of the computer room air conditioner (CRAC) blower speed (N_{CRAC}), the CRAC return air temperature set-point ($T_{a,ret}$), and the information technology (IT) load distribution for the racks ($\dot{Q}_{room}$; where, $\dot{Q}_{room} = \sum \dot{Q}_{racks}$), while rack inlet air temperature ($T_{\text{rack inlet}}$) was the predicted variable. The transient model, which considers a cooling failure scenario, predicts the rack inlet temperature ($T_{\text{rack inlet}}$) as a function of time (t), which is the only independent input parameter.

The performance of the four data driven models was evaluated based on the absolute mean error for interpolation and extrapolation, and the adaptability of the models to changes in the physical domain (data center room) configuration. For the steady-state case study, the predictions for ANN, SVR and GPR models were in good agreement with CFD/HT simulations, with the GPR model having the smallest overall average prediction error of 0.6 °C in rack inlet air temperature, corresponding to a relative error of 2.7% with respect to rack inlet temperature measured in °C. Interestingly, the prediction error in rack inlet temperatures for all the models was found to increase with altitude, *i.e.*, vertical distance from the floor. This was attributed to under-provisioning and recirculation effects that are not entirely captured by these models. For the transient case study, the interpolative prediction error for all the models is very low (< 0.3 °C); however, the extrapolative prediction errors were much greater, and appear to be directly proportional to the (here, temporal) "distance" from the interrogation point to the input parameter space.

In data driven modeling, generating the training data set is the most time-consuming component. It is therefore critical to determine the minimum size of the training data set required to develop a model with reasonable fidelity. As such, the impact of the size of the training data set on prediction accuracy was also compared for the four models. The GPR model was found to have the best accuracy for smaller training data sets compared with the other models, with an average prediction error for rack inlet temperatures < 1 °C when trained on only 50 simulations.

Another contribution of this dissertation is the development of a cooling energy estimation model as detailed in Chapter 5. Thermodynamic analysis of various components

was carried out to develop a relationship between the operational state of a data center and overall power consumed by cooling infrastructure corresponding to the operational state.

Finally, Chapter 6 presents the development of a GA-based static and quasi-static optimization method and the integration of all the aforementioned components (DDM, cooling energy estimation model and optimization model) in one holistic framework. The static optimization problem considered informs IT load distribution and cooling set-points in the data center room to simultaneously optimize two objectives: 1) minimize cooling power consumption, and 2) maximize IT load. Three optimization scenarios, Room-Level IT Load, Row-Level IT Load Distribution and Rack-Level IT Load Distribution, that consider IT load distributions at different spatial resolutions, are considered. Results for all three scenarios in the form of non-dominant solution series, *i.e.*, the Pareto-Front (PF), were presented. Tests were conducted to compare the minimum cooling power required corresponding to the three IT load distribution scenarios when the overall load in the data center is varied from 40 kW to 160 kW. For any given IT load, the results suggest that optimization employing rack-level IT load distribution minimizes cooling power.

A quasi-static framework that aims to optimize cooling power consumption in the data center during operation was also developed. For a given IT workload distribution, the framework determines most energy efficient set-points for the cooling infrastructure while preventing temperature overshoots. The framework was implemented for a test run of 7.5 hours with a step change in room IT load every 30 minutes. The results are compared to a baseline operation case where the CRAC set-points are kept fixed with active CRAC return air temperature control and a case where the set-points are constant with no active CRAC return air temperature control. The results indicate that the developed framework can

reduce cooling energy consumption by 18% compared to the case with constant cooling set-points with return air temperature control, and by 52% compared to the case with constant cooling set-points and no return air temperature control. For the Data Center Laboratory at Georgia Tech a reduction in cooling power consumption by 18% annually would translate to a saving of approximately $7000 in electricity cost per year.

Depending on their physical size, data centers are classified into many categories ranging from closet level to hyperscale data centers [10]. Based on this classification, the Data Center Laboratory at Georgia Tech would be considered a 'Server-Room'. Consider instead a mid-tier data center (size: 2000-19,999 ft^2) with 100 racks, each with a power density of 12 kW, and thus a total IT power consumption of 1200 kW for the entire facility. According to The United States Data Center Energy Usage Report [10], the average cooling PUE (equation 7.1) mid-tier data centers $PUE_{cooling}$ = 0.9, where:

$$PUE_{cooling} \equiv \frac{\text{Cooling Energy Consumption}}{\text{IT Equipment Energy}} \tag{7.1}$$

Assuming that the facility runs continuously with an average utilization of 70% the annual cooling energy consumption for this mid-tier data center would be ~6.7 million kWh. Implementing the developed optimization framework and assuming that only a 10% reduction in cooling power consumption can be realized, the saving in electricity costs for this facility would be $80,000 per annum. Note that the savings in electricity costs would be even greater for larger enterprise and hyperscale data centers.

7.1 Implications and Discussion

This thesis presented an overall framework for cooling energy optimization in data centers (see Fig. 62) and demonstrated its application to the Data Center Laboratory at Georgia Tech. This framework, with some modifications, can be employed to optimize operation and minimize cooling energy consumption for other facilities. This section details how each component of the developed framework can be adapted when considering a new facility and factors that would need to be considered.

7.1.1 Data Generation

In this work training data for data driven modeling were generated by conducting parametric CFD simulations. Depending on the configuration of facility under consideration, the availability of external and on-board sensors in the data center and the number of input/output parameters considered for modeling, training data can be generated instead by conducting parametric experimental tests or a combination of CFD/HT simulations and experimental measurements.

7.1.2 Data Driven Modeling (Thermal Model)

This work compared the performance of four data driven models for steady state and transient prediction. As mentioned in Chapter 4, each of the modeling frameworks offer specific advantages and an appropriate model can be selected based on the data center size/configuration as well as the size of the available training dataset.

Another important aspect to be considered is the desired/acceptable accuracy of the data driven model developed. As indicated in static and quasi-static optimization formulations (Figs. 54 and 58), the average error in rack inlet temperature prediction ($\varepsilon_{prediction}$) is accounted for when enforcing the prescribed temperature threshold as a constraint function. Thus, for a greater prediction error in rack inlet temperature, a more conservative cooling set-point will be determined by the optimization framework to satisfy the constraints. This reduces the energy savings that could potentially be achieved compared with the case when a thermal model with higher prediction accuracy is used. It has been estimated that for every 0.5 ℃ decrease in temperature set-point, the cooling power consumption increases by 1-2.5%. For the mid-tier data center facility (100 racks) described earlier, a conservative temperature set-point would thus increase electricity costs by $20,000 each year. Moreover, this amount does not include the increase in the electricity consumed by the blower, which is proportional to the cube of the blower speed set-point. These factors will both significantly increase the operation costs for a data center facility.

One of the challenges that needs to be considered when employing data driven models for thermal modelling in data centers is the scalability of a developed model to a larger facility with a similar physical configuration. An approach to tackle this could be to construct an appropriate zonal model which acts as a building block. Multiple such zonal models could be "stitched" together to construct a model for a large facility. An important aspect would be treatment of interfaces between adjacent zonal models. One possible option for interface treatment would be to experimentally measure boundary conditions at the interface and include them as parameters in the input training data.

7.1.3 Cooling Power Consumption Model

The cooling power consumption model presented in Chapter 5 takes into account specific details of cooling system at the Data Center Laboratory at Georgia Tech. However, models to estimate cooling power can be developed for any air-cooled facility by tracing heat flow from data center room (source) to the ambient (sink) and conducting thermodynamic analysis for the involved components. When available, experimental measurements for power consumed can be used to validate the model developed.

7.1.4 Optimization Framework

Formulation of optimization problem is closely linked to the objective function (here, cooling power consumption model) and constraint function (here, thermal model for rack inlet air temperature prediction). In case a data driven thermal model is employed, the number of optimization variables considered in a GA based optimization would be less than or equal to number of input parameters considered while training the model. The number of optimization variables in turn would dictate the ideal population size for every generation as well as the computation time required to solve the optimization problem.

7.2 Recommendations and Future Work

This thesis has demonstrated an overall framework for cooling energy optimization in data centers. This research has suggested a few complementary research areas as well as topics whose further study could enhance the impact of this work as described below:

- The quasi-static optimization framework in this research employed the return air temperature as a cooling set-point (and optimization variable), largely due to constraints posed by design of the CRAC units in the data center lab. It may therefore be worth considering the CRAC supply air temperature or rack inlet temperature itself as a cooling set-point (optimization variable). This would require configuration changes, including replacing the feedback logic of the on-board controller (PID), for the CRAC unit.

- The adaptability of data driven models to physical configuration changes is limited due to assumptions inherent to the modelling framework. This in turn limits the generalized applicability of a given model (in its original form) across data centers with different configurations. An approach which would include physical configuration attributes, *e.g.* the number of aisles in the data center, number of racks in an aisle, type of servers, relative position of cooling units etc., as input parameters for training the data driven models could potentially address this issue to a certain extent. Data can be collected from a large number of data centers with varying configurations and used to train data driven models.

- The framework developed and presented in this dissertation aims at controlling cooling infrastructure to minimize cooling energy consumption in data center operation. Another approach to minimize energy usage is the development of "thermally-aware" frameworks for IT workload distribution and migration. Combining these two approaches into an

integrated optimization platform could further improve the energy efficiency of data centers.

- Free air cooling /outside air economization should be considered and integrated with the developed framework to minimize cooling power consumed by the chiller unit in regions with suitable weather conditions.

APPENDIX A: DETAILS OF LAB EQUIPMENT AND MEASUREMENT TOOLS

A.1 Lab Equipment Configuration:

Table A1: Equipment Configuration

Equipment	***Details***
PDU	Liebert PPA225C, Capacity-225kVA
CRAC	Liebert FH740C, Capacity- 60 Ton
Rack Configuration -	
(a) A1	Networking Cabinet
(b) A2	IBM Blade Center, Capacity – 30.9 kW
(c) A3	Server Simulator, Capacity – 21 kW
(d) A4	IBM xSeries 335, Capacity - 14.35kW
(e) A5	HP Proliant 360, Capacity – 16.35 kW
(f) B1	Dell PowerEdge 2850, Half filled (11 servers), Capacity – 7.7kW
(g) B2	IBM xSeries 335, Capacity – 14.35 kW
(h) B3	IBM xSeries 335, Capacity – 14.35 kW
(i) B4	HP Proliant 360, Capacity - 18.45 kW
(j) B5	Dell PowerEdge SC1435, Capacity – 25.8 kW

A.2 Experimental Measurement Tools

A.2.1 Tile Airflow Rate Measurement Tool

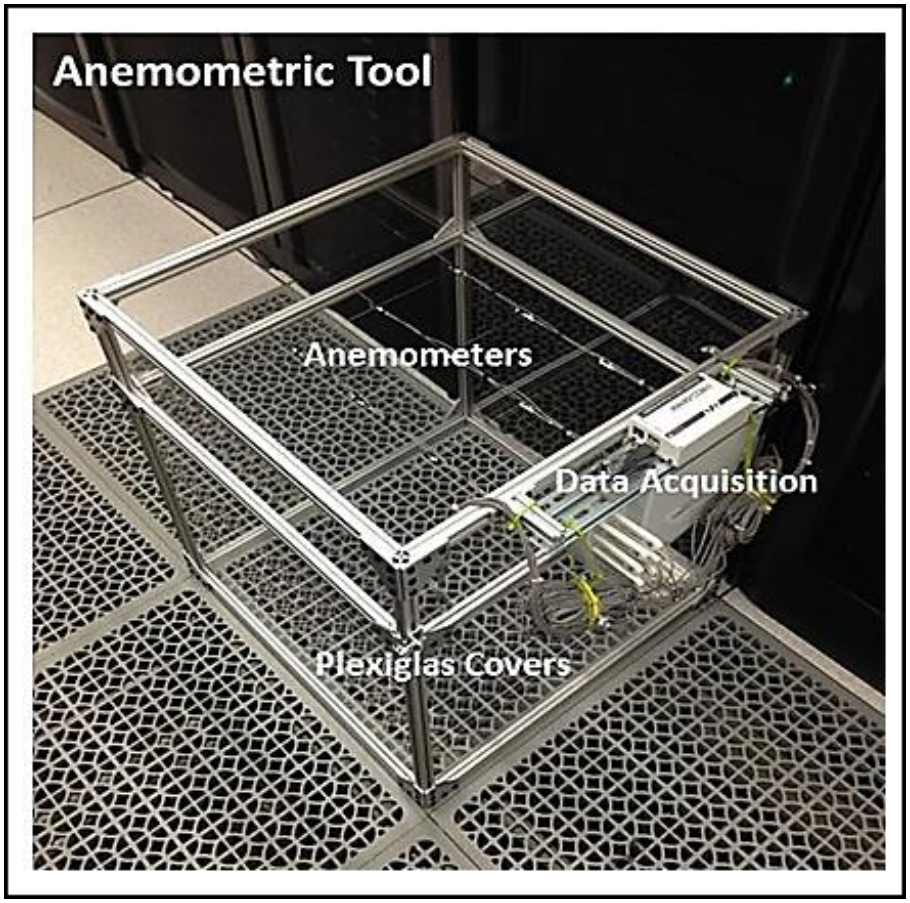


Figure A1 - Array of 16 Anemometers for Tile Air Flow Rate Measurement

A.2.2 Rack Airflow Rate Measurement Tool

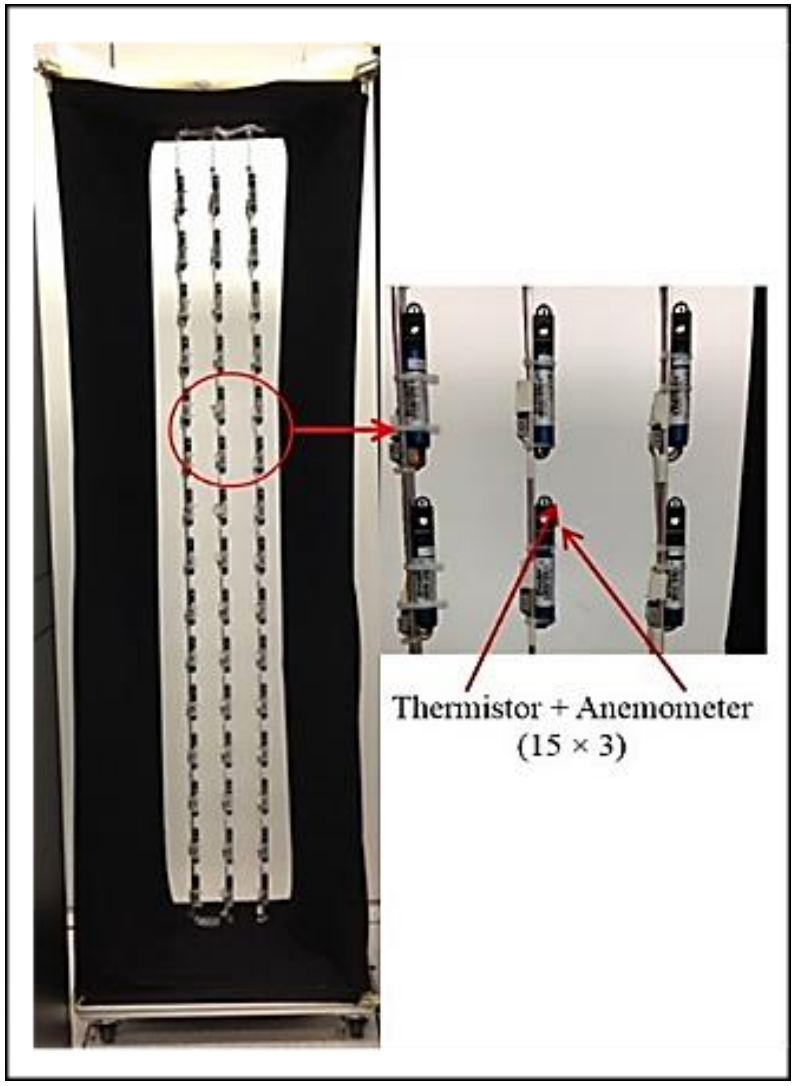


Figure A2 - Array of Anemometers for Rack Air Flow Measurement

A.2.3 Rack Inlet Air Temperature Measurement Tool

Figure A3 - Rack Inlet Air Temperature Measurement

APPENDIX B: IMPACT OF ACTIVE TILES ON DATA CENTER FLOW AND TEMPERATURE DISTRIBUTION

Localized hot spots pose a major problem for data center thermal management. At present, most centers respond to hot spots by increasing the server fan speed to provide more cold air at the location of the hot spot. In an attempt to make sure that each server operates under temperature thresholds many servers are over-cooled. Thus, ensuring that cold air is available everywhere at all times requires over-provisioning the data center. Conversely, a lack of cold air causes recirculation of warm room air, making the hot spot worse. The desirable approach is instead to provide additional cold air only where and when required, and avoid overcooling the entire data center.

Active tiles with integrated fans increase the local volume flowrate by redistributing the cold air supplied by the CRAC (computer room air conditioning) to the underfloor plenum. The objective of this work is to determine whether active tiles can be used to efficiently and economically eliminate hot spots in data centers. Experiments were conducted to determine air flow and temperature distributions, as a function of the CRAC unit blower and active tile fan speeds, for both a single tile/rack combination and a complete aisle populated with active tiles. The physical configuration of the data center room is described in Chapter 2 (Figs. 10 and 11). The results were compared with those for passive (generic) tiles with similar effective porosity. Cross-correlation factors representing the effect of active tile on adjacent tiles are presented, and the power consumed by the active tiles in terms of W/m^3 air at the tile outlet is compared with that

Reference: J. Athavale, Y. Joshi, M. Yoda, and W. Phelps, "Impact of Active Tiles on Data Center Flow and Temperature Distribution," in *15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 1162-1171.(2016)

for passive tiles. The results indicate that the power consumption per unit air flow can be reduced by using a full aisle of active tiles.

Based on the experimental observations, a control framework that uses the active tile fan speed as a process variable to provide the air flowrate required to cool the hot spot, while reducing power consumption by the server fan, and hence the energy consumed by the data center can be designed.

B.1 Introduction

Thermal management of data centers remains a challenge because of ever-increasing power density and ever-smaller server footprints. These trends mean that the fraction of total power required to cool a data center and ensure that all the servers remain within the prescribed temperature limits, continues to increase. Indeed, as much as 50% of the total energy consumed by data centers at present is used for cooling. It is therefore essential to develop improved cooling schemes that can reduce energy consumption without compromising server reliability.

Figure B1 illustrates the different cold air flow paths used to cool the IT equipment in the data center. The tile and server air flow rates are related to the different data center parameters as described in equations (B.1) and (B.2).

$$\dot{V}_{tile} = f(N_{CRAC}, Spatial\ Location, Tile\ Configuration) \tag{B.1}$$

$$\dot{V}_{rack} = f(N_{server}) = f(T_{rack\ inlet}, T_{CPU}, P_{rack}) \tag{B.2}$$

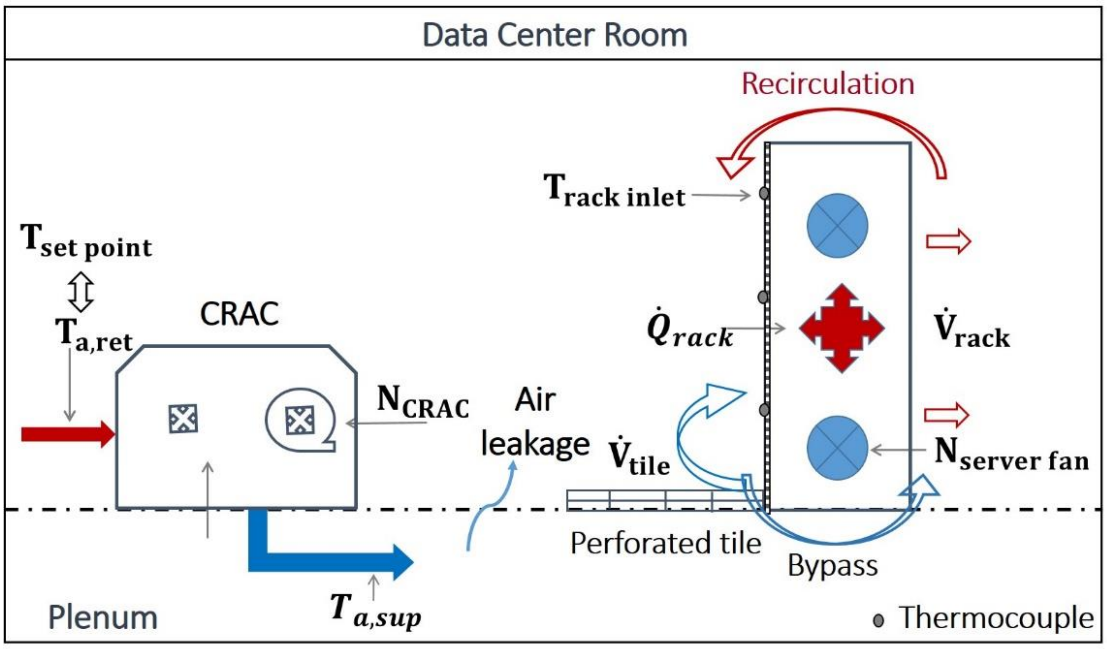


Figure B1 – Data Center Operation

An imbalance between these two flow rates leads to undesirable phenomena such as recirculation, bypass and leakage, as depicted in the Fig. B1. The global conditions of the data center can be altered by manipulating the CRAC set-point temperature, mass flow rate of cold air being supplied to the room, and overall heat load for the room. The local conditions in a particular section of the rack depend on the amount of cold air available at that location, corresponding server fan speed and power consumption for the particular server or group of servers. Controlling both the global and local conditions is required to ensure reliable operation, and hence the cooling strategy should be optimized and controlled at both levels.

The idea of dynamically and locally allocating cooling resources as required was suggested and implemented by Patel et al. [55, 56]. They computed regions of influence for the CRAC units within a data center, and modulated the set-points for the units based

on local data gathered from the corresponding regions resulting in lower energy consumption and improved use of the critical data center space. Although this control architecture ensured that the cooling needs of smaller sections within a data center were sensed and met individually, it was inefficient at even smaller, *i.e.*, the individual rack or server, scales.

There are two requirements to optimally provision individual servers or a group of few servers with cold air: first, there must be enough cold air available, and second, this cold air delivered by the floor tiles must be drawn in by the server fans at the location of the servers. Traditionally, server fans have been designed so that they have a sharp step response in their rotation rate when the front face temperature exceeds a preset threshold value [2, 129]. Such a coarse discrete response is neither optimal in terms of energy efficiency nor providing adjustable cooling. Wang et al. [51] designed a multi input multi output MIMO controller to proactively tune the server fan speed based on server temperature. Their results indicate that using optimal control can give finer and more sensitive temperature control, while decreasing the server fan power consumption by 20% compared with standard server fan control.

Ensuring that cold air is available when and where it is required without overcooling the whole data center space requires controlling the distribution and supply of cold air from the CRAC units. Boucher et al. [61] conducted experiments varying parameters such as the CRAC supply temperature, blower speed and tile vent opening to determine their effects on data center behavior. They reported that varying the opening and relative position of the tile vents had a predictable effect on the local rack inlet temperature and could therefore potentially be used for local cooling control. A control algorithm based

on modulating the vent tile openings was designed and implemented by Beitelman et al. [49]. An online model estimator and adaptive MIMO controller were used to estimate the rack inlet temperature, and change the vent openings to maintain the inlet temperature below a prescribed threshold. The results demonstrated that adaptive vent tiles could be used to fine tune the local temperature and optimize air flow distribution.

Active floor tiles are porous, with integrated fans, as shown in Fig. B2 [77]. In order to understand how and whether they can be used to dynamically and locally allocate cooling in a data center, experiments were performed on a single active tile and rack, as well as a complete cold aisle populated by active tiles. Additional experiments to quantify cross-correlation factors for active tiles were also conducted and the results are presented below.

Figure B2 – Active Tile

B.2 Active Tile Porosity Characterization

Before comparing our results with those for passive tiles of a similar porosity, experiments were conducted to determine the effective porosity of active tiles. For the

turbulent flows characteristic of data centers ($Re > 10^4$), the pressure loss across a restriction is proportional to the specific kinetic energy of the air flow; the constant of proportionality is called the pressure loss factor ($\Delta p = K(\rho V^2 / 2)$) [9, 10].The difference between the plenum and room pressures can then be written in terms of this loss coefficient and the volumetric flow rate through the tile [130, 131].

$$p_p - p_r = \left(\frac{\rho}{2A^2} \right) K_{coeff\ t} \dot{V}_t^2 \tag{B.3}$$

Assuming that the pressure difference between the plenum and the room pressures is the same for both the passive tile and unpowered active tile, we obtain:

$$K_{coeff\ a} = \frac{K_{coeff\ p}}{\left(\frac{\dot{V}_a}{\dot{V}_p} \right)^2} \tag{B.4}$$

where the subscripts *p* and *a* denote passive and active tiles, respectively. The flow rates through both types of tiles were obtained by consecutive measurements at the same location by switching tiles; the fans on the active tile fans were off during the measurement. We compute $K_{coeff\ a}$ using the manufacturer's value for $K_{coeff\ p}$ and measured tile flow rates in (B.4). The effective porosity for active tiles can then be calculated using the following equation [130]:

$$K_{coeff\ tile} = \frac{1}{F^2}\left(\left(\frac{1-F}{2}\right)^{\frac{1}{2}} + (1-F)\right)^2 \qquad \text{(B.5)}$$

These calculations gave an active tile porosity of 27%. This porosity was confirmed by several measurements obtained for a range of CRAC blower speeds and tile locations.

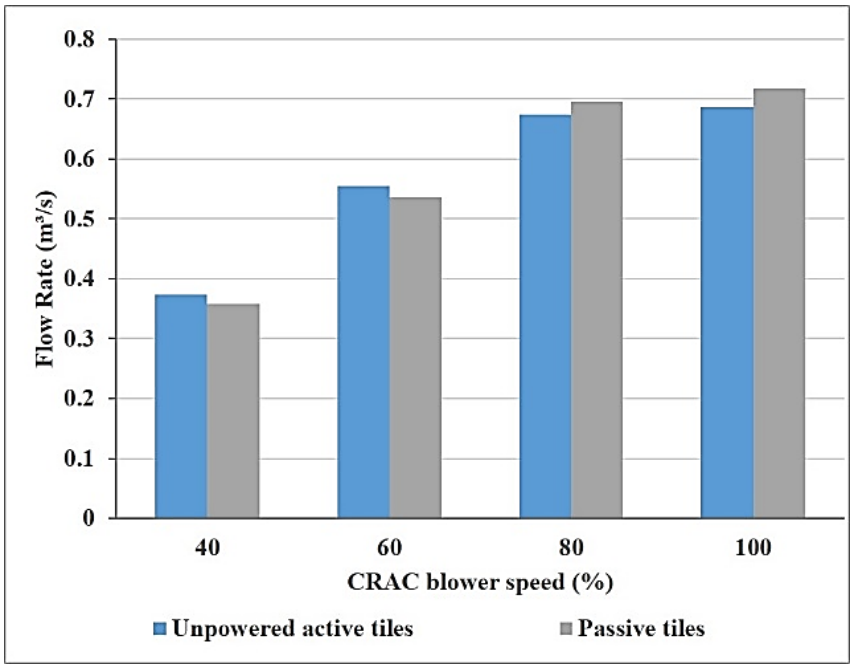


Figure B3 - Tile Flow Rate Comparison for Active Tile and Passive Tile with Porosity =27%

Figure B3 shows the tile flow rate for passive and unpowered active tiles having the same porosity and measurement location as a function of CRAC blower speed. The tile flow rate is approximately the same for both types of tiles. All the passive tiles considered in the rest of this study therefore have a porosity of 27%.

B.3 Experimental Results and Discussion

B.3.1 Results for Single Active Tile and Rack Configuration

For the experiments in this section, only one (out of a total of ten) was an active tile, and the rest were passive tiles. A server simulator was used to model a rack next to the active tile since the server fan speed and thus the rack air flow rate should remain constant throughout the experiments. Figure B4 shows the relative location of the active tile and server simulator in the room.

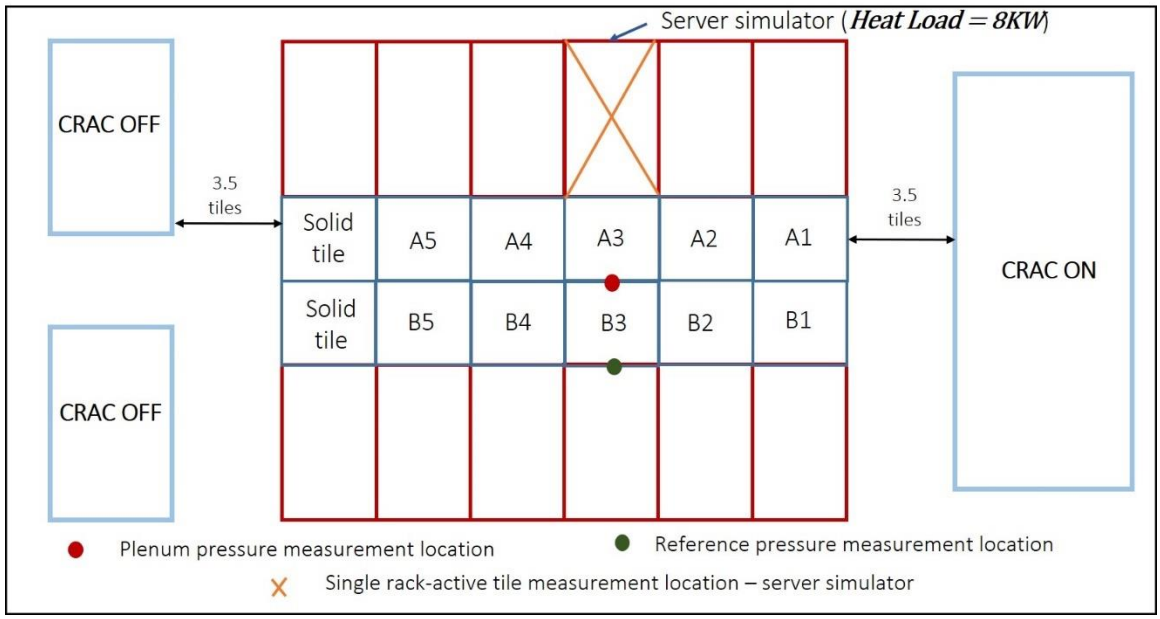


Figure B4 – Location of Server Simulator and Active Tile (A3)

To non-intrusively control the speed of active tiles, the active tiles were programmed to respond to a voltage input provided by National Instruments NI-FP1601 controller and a voltage power supply unit. The communication signals to the controller were sent via wireless network so that the active tile fan speed could be controlled from a location outside the data center room. The voltage output of the NI-FP1601 is connected to the SH-200 terminal which communicates directly with the active tiles. The active tile

speed varies from 10% to100% of the maximum rotational speed as the voltage signal varies from 1 to 10V. Only one CRAC unit was used in these experiments, as shown in Fig. B4. For these experiments CRAC supply temperature and server simulator heat load were kept constant at 15.5℃ and 8 kW respectively.

Three different experiments were conducted to study the effect of a single active tile as follows:

B.3.1.1 Tile Air Flow Rate Results

In the first set of experiments, the tile flow rate was measured as a function of the CRAC blower speed and active tile fan speed, which were varied independently. A grid of 16 thermal anemometers, as shown in Fig. A1 in Appendix A, was used to measure the air velocity. The air flow rate was then the product of the grid area and the average velocity over these 16 readings [78].The tool has an uncertainty of+-5% for air velocity measurement as specified by the manufacturer and confirmed using hand held anemometer. Figure B5 shows the flow rate data as a function of the active tile fan speed, where different curves correspond to different CRAC blower speeds. In this figure, an active tile fan speed of 0% corresponds to the case of passive tiles, which are used as the baseline to estimate the increase in tile flow rate. The tile flow rate increases monotonically with active tile fan speed, as well as with CRAC blower speed. However, the actual increase in the tile flow rate over the baseline case decreases as the CRAC speed is increased. Indeed, the curve at the lowest CRAC blower speed has the greatest increment in flow rate (along Y axis) as active tile speed goes from 0 to 100%. This behavior can be explained by the fact that it is difficult for the relatively small active fans to modify the high plenum pressures which

appear at high CRAC blower speeds. This suggests that active tiles will have the greatest effect in regions of low plenum pressure areas or at low CRAC blower speeds.

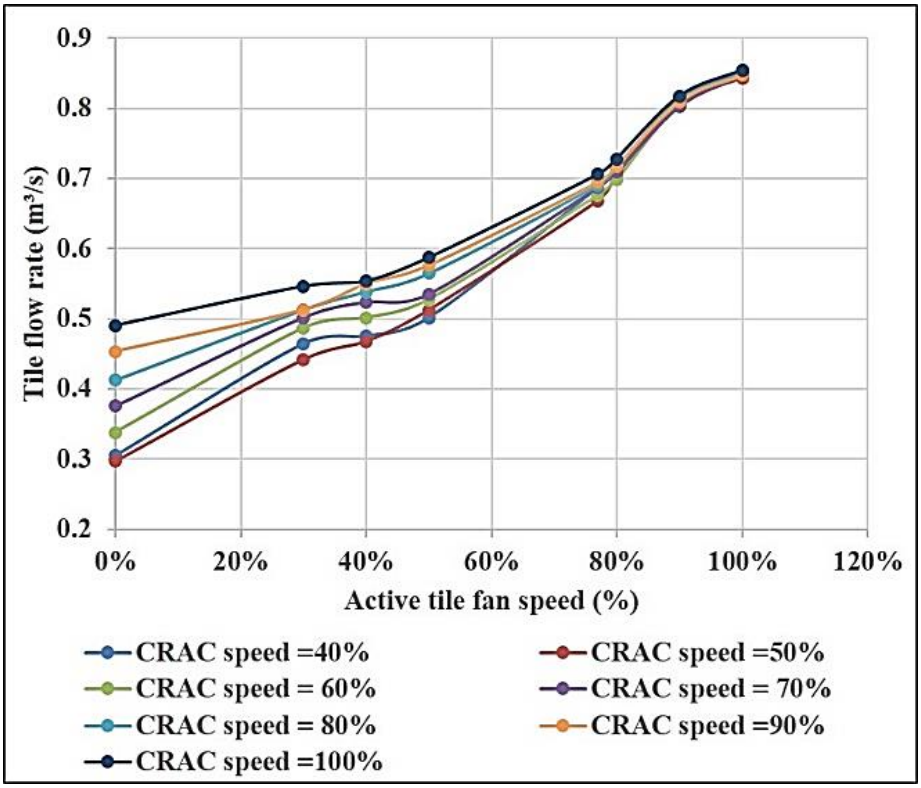


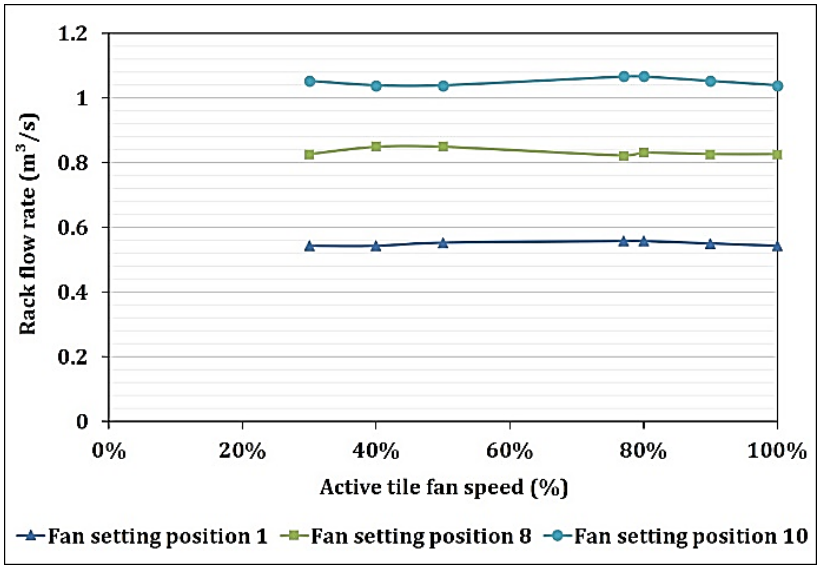
Figure B5 - Tile Air Flow Rate vs. Active Tile Fan Speed for Different CRAC Blower Speed

B.3.1.2 Rack Air Flow Rate Results

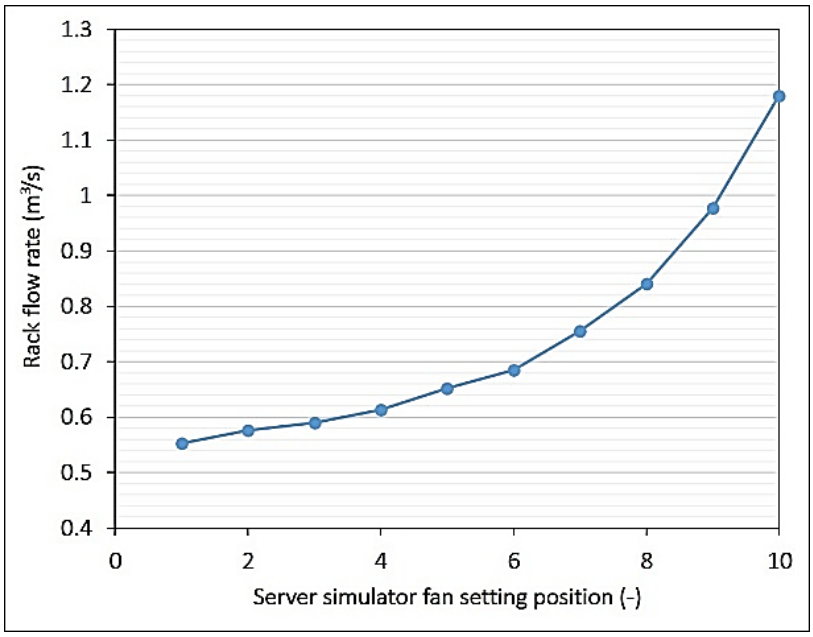
As stated previously, the flow rate through a rack depends only on the server fan speed. This dependence was validated by measuring and comparing the flow rates through the rack for two cases: 1) Varying the active tile fan speed while keeping the server fan speed constant, thereby supplying increasing amounts of air; and 2) Increasing the server fan speed while keeping a constant active tile fan speed. The device used to measure the rack flow rate, shown in Fig. A2 [77] in Appendix A, consists of an array of 45 thermal anemometers perpendicular to the flow placed at the back of the rack, with a cloth shield

to direct the flow through the sensor array. The measurements of rack flow rate have an uncertainty of ±10% which has been determined by making calorimetric measurements.

As seen in Figs. B6 (a) and B6 (b)the rack flow rate is a function of server fan speed only and is independent of the active tile fan speed and therefore the amount of air supplied. So, for an under-provisioned data center the deficit in cold air would lead to the server fans drawing in hot air from the room. However, supplying additional cold air to maintain appropriate server temperatures will not be useful if the server fans do not run at a speed that would draw in this air. This demonstrates the importance of supplying exactly the required amount of cold air to all the locations in a data center.



(a)



(b)

Figure B6- (a) Rack Flow Rate vs. Active Tile Fan Speed for Server Fan Setting (b) Rack Flow Rate vs Server Fan Setting

B.3.1.3 Rack Inlet Temperature

The third set of experiments on the single active tile configuration measured the temperature distribution over the front face of the rack. The internal fan setting for the

server simulator corresponded to an air flow rate from the data center room of $0.67 m^3/s$. Three different scenarios were considered, corresponding to whether the server simulator is under, correctly, or over-provisioned using a single active tile. The results were compared with a baseline case using only passive tiles, as summarized in Table B1.

Table B1: Test Cases for Rack Inlet Temperature Study

Case Number	Case	Setting	Flow Rate (m^3/s)	Remarks
1.	Under-provisioned	Active tile fan speed =40%	0.50	25% less than required
2.	Exactly-provisioned	Active tile fan speed =77%	0.67	~ required flow rate
3.	Over-provisioned	Active tile fan speed =100%	0.85	25% more than required
4.	Baseline	Passive	0.33	-

Figure B7 presents temperature distributions and area-averaged temperatures over the front face of the server simulator for these four cases. The temperature profiles were obtained using a grid of 108 T-type thermocouples with a measurement uncertainty of $\pm 0.5℃$ and bilinear spatial interpolation between the measurement points. The average temperature over the front face decreases as the active tile fan speed increases. In the case with only passive tiles and the under-provisioned case there are high-temperature regions at the top of the rack that are absent in the cases when the server simulator is exactly and over provisioned using the active tile. From the tile flow rate measurements, the highest flow rate that can be obtained using passive tiles is $0.49\ m^3/s$ when the CRAC is running at 100% (see figure B5), which is a much lower flow rate than that required by the rack in this instance. Table B2 presents a comparison of power consumption for the cases

examined and the case using only passive tiles for the CRAC running at 100%. Note that total power consumption in case of active tile includes the power consumed by the tile in addition to the power consumed by the CRAC unit.

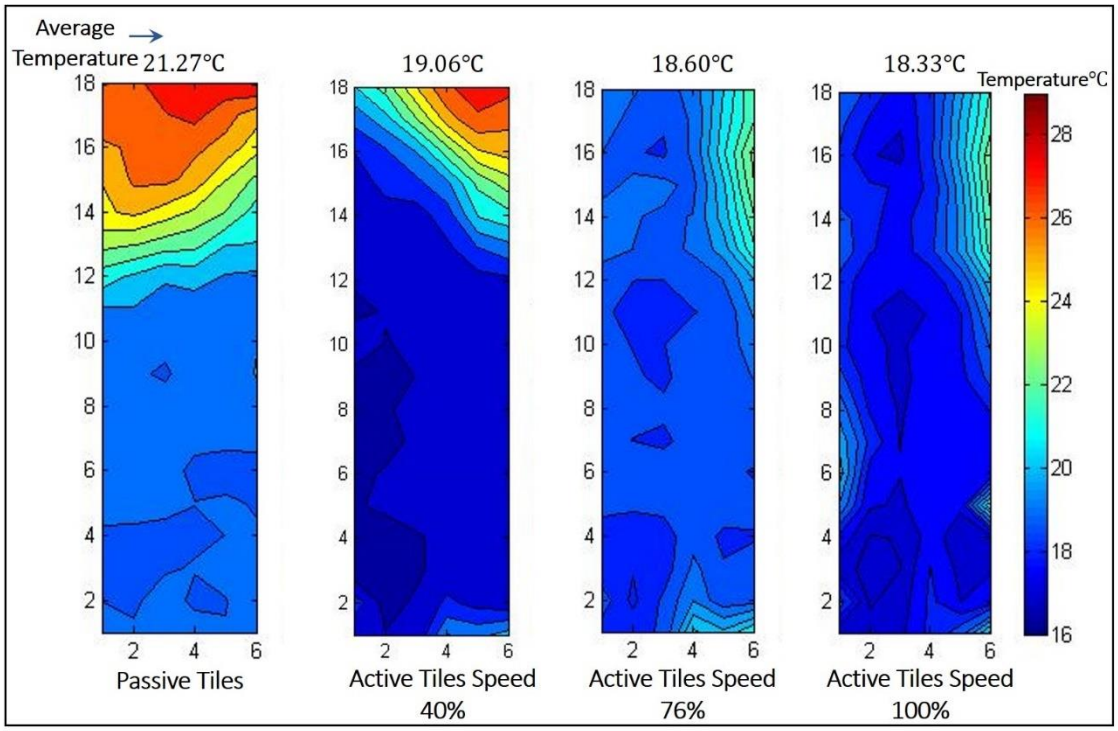


Figure B7 - Front face and Area Average Temperature for Server Simulator

Table B2: Test Cases for Power Consumption Comparison

Case Number	Case	Setting	Total Power Consumption (W)
1.	Under-provisioned	Active tile fan speed =40% CRAC speed=60%	1,004
2.	Exactly-provisioned	Active tile fan speed =77% CRAC speed =60%	1,041
3.	Over-provisioned	Active tile fan speed =100% CRAC speed =60%	1,094
4.	Baseline	Passive CRAC speed =60%	950
5.	Passive (max flow rate)	CRAC speed =100%	3,000

Although the power consumed in case 5 is much higher than that of the exactly-provisioned case, the passive tiles still cannot provide the cold air flow rate required to cool the server This suggests that a rack with high heat load can be supplied with the required air flow rate much more efficiently using a combination of an active tile and CRAC unit. Since CRAC blowers supply air to the entire data center, increasing the blower speed would increase the flow rate at all the tile locations, instead of at the specific servers with high heat loads. Increasing the blower speed will therefore likely over-cool the regions with lower heat loads, while under-cooling the regions with higher heat loads.

B.3.2 Results for Complete Aisle of Active Tiles

This section describes and compares experiments on an entire aisle of tiles, both active and passive. The data center space is the same as that for the previous set of experiments, except that all ten tiles are either active or passive. The tiles are numbered as shown in Fig. B4. Only one tile is directly connected to the controller; the remaining 9 are connected serially to control their fan speeds. The voltage control for changing the tile fan speed is implemented in pairs (*e.g.* Tiles (A1&B1), (A2&B2), etc.) to ensure that both the tiles in a pair will have the same fan speed independent of the other (pairs of) tiles. Only one CRAC unit was used in the following experiments, as shown in Fig. B4, and the CRAC supply temperature was set to 15.5℃. Three different studies were conducted for the entire aisle of tiles as follows.

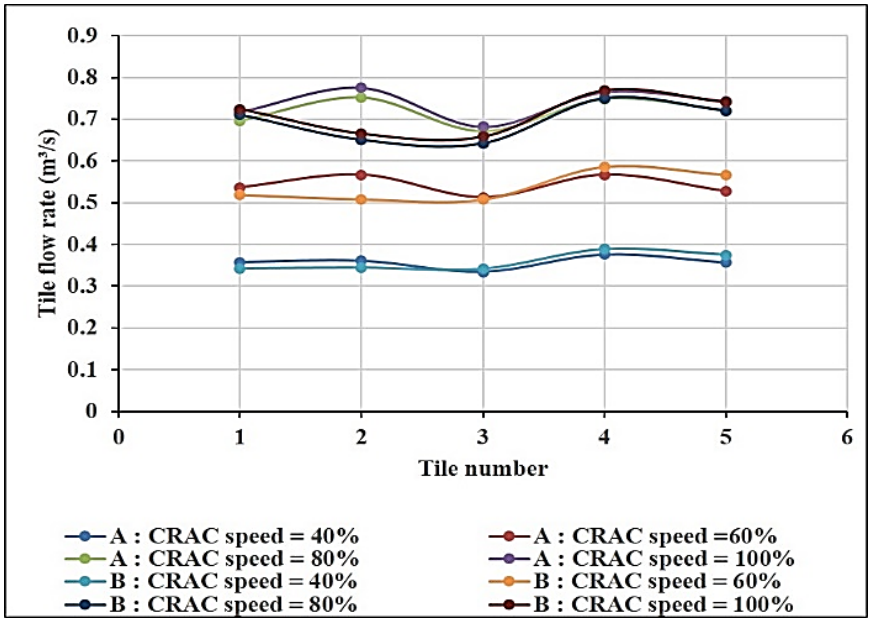
B.3.2.1 Tile Air Flow Rate and Plenum Pressure Measurements

The flow rate through individual tiles was examined as a function of CRAC blower speed and the active tile fan speed. The tile flow rates, measured by the thermal

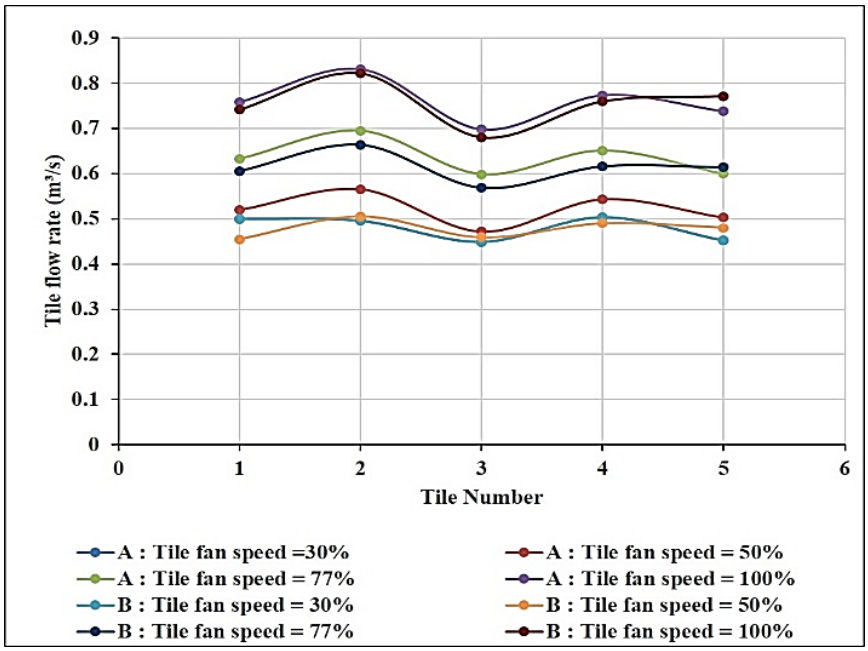
anemometer array (see Fig. A1) are presented in Fig. B8. Figure B8(a) shows tile flow rate as a function of tile location at different CRAC blower speeds for the case where the aisle is populated by passive tiles. Figure B8 (b) shows a similar plot for the case where the aisle is populated by all active tiles, except that the different curves now represent different active tile fan speeds and the CRAC blower speed is set at a constant 40% of maximum. The figures show that the tile flow is a unique function of the tile location in the data center space, and that the flow rate through a tile increases with CRAC blower speed, or the active tile fan speed. The flow rate through the active tiles is slightly higher than that for the passive tiles.

Figure B9 shows the plenum static pressure compared with the data center room pressure. The measurements were made using an Alnor micro manometer AXD610 [79] with an accuracy of $\pm 1\%$ as indicated by the manufacturer. The manometer was connected to the plenum pressure measurement location and reference location using 4.8 mm internal diameter PVC tubes. The plenum pressure was measured in the mid plane of the cold aisle, and 0.3 m (1 ft.) below the tiles. To measure the room pressure, the tube was attached at the top of rack (see Fig. B4). The tubes are oriented facing upwards which is the direction of airflow in the plenum. From the graph it is seen that the relative plenum pressure increases with an increase in the CRAC blower speed, as more and more air is pumped into the plenum. For a given CRAC blower speed the relative plenum pressure decreases as the active tile fan speed is increased. This is expected since the active tile fans are connected in series with the CRAC blower, resulting in a decrease in the intermediate pressure between the two. The lower plenum pressure in case of using active tiles also suggests that there is minimal air leakage from the plenum to the data center room, with almost all the

cold air exiting through the perforated tiles. This is one of the primary reasons for overall increase in tile flow rate when using active tiles. The ideal case would be to have relative pressure to be positive but approaching zero since this would result in minimum air leakage from the plenum to the data center room.



(a)



(b)

Figure B8 - (a) Tile Flow Rate vs. Tile Location for Different CRAC Blower Speeds (b) Tile Flow Rate vs. Tile Location for Different Active Tile Fan Speed

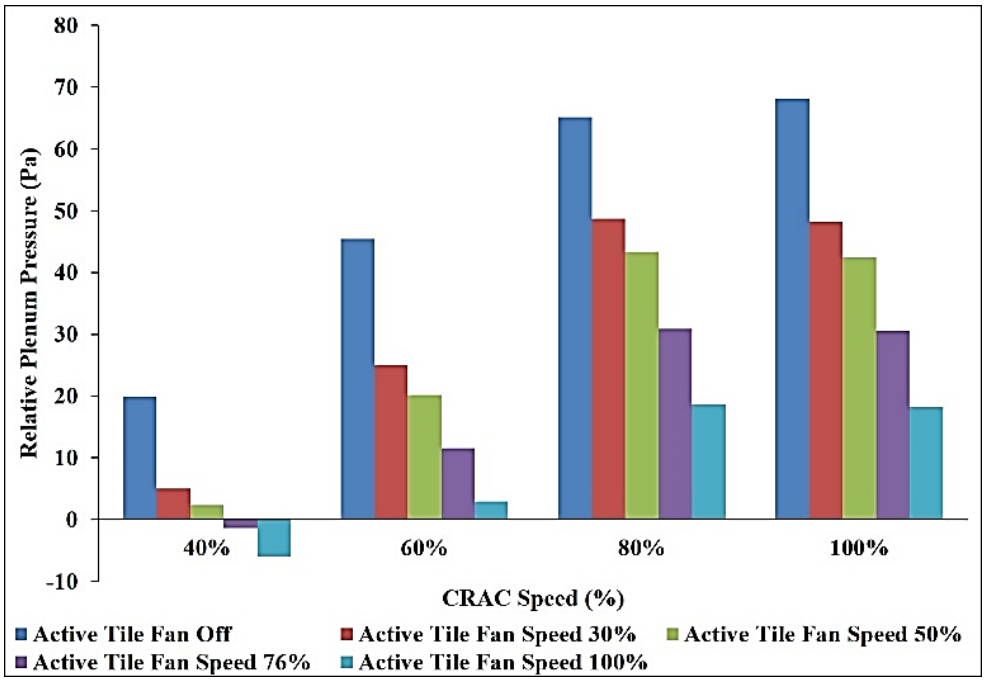


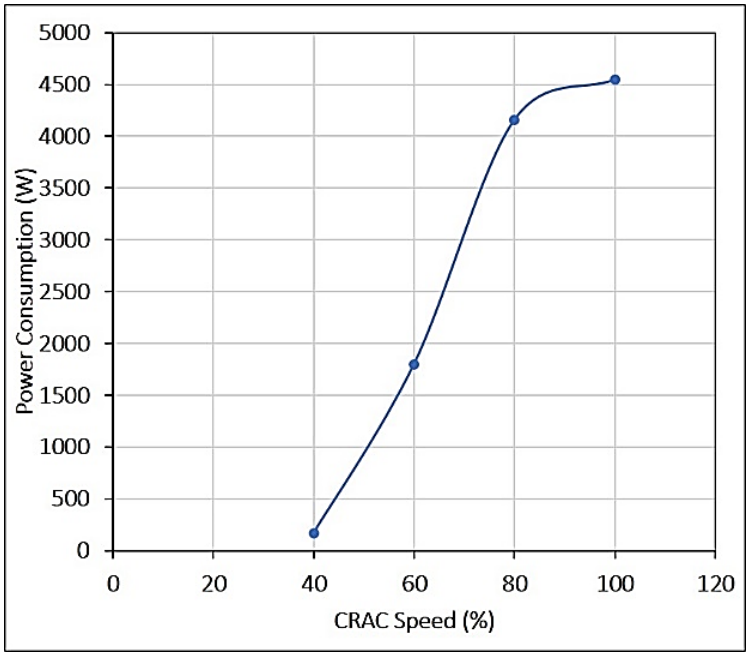
Figure B9 - Relative Plenum Pressure as a Function of CRAC Blower Speed and Active Tile Fan Speed

B.3.2.2 Power Consumption Results

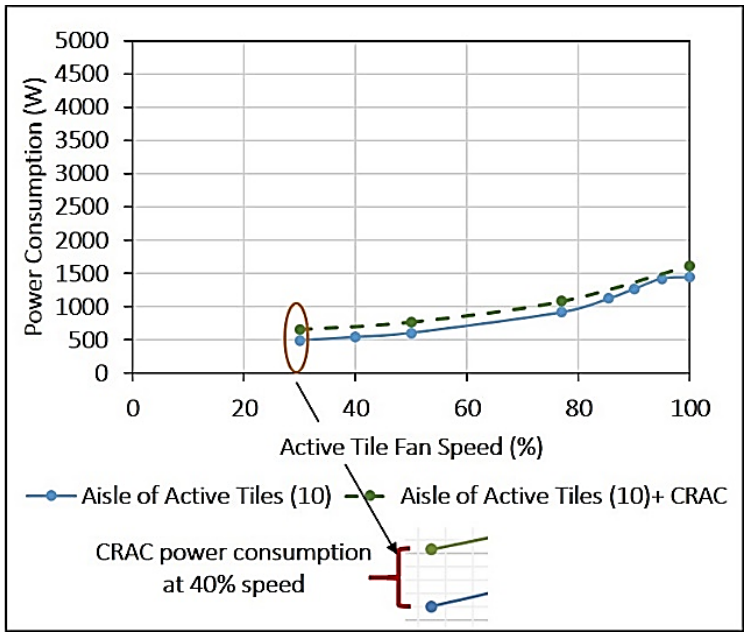
Energy consumption is an important metric for evaluating the performance of data centers. Usually, the objective in data center operation is to minimize energy consumption subject to the constraints prescribed by the thermal management thresholds. It is therefore important to measure and reduce, if possible, the cooling power required to maintain the appropriate server temperatures. The total power consumption for the two tile configurations was measured here. First, the CRAC speed, and hence the total flow rate is varied for the case where the aisle is populated by all passive tiles. Here, the total power consumption is taken to be the total power as reported by the CRAC unit. The total flow rate is the sum of the flow rates through all ten tiles. Second, the active tile fan speed is varied to change the flow rate through the tiles for the case where the aisle is populated by all active tiles at a constant CRAC blower speed of 40% maximum. Although the total flow

rate is still the sum of flow through all ten tiles, the power consumption in this case includes both the power reported by the CRAC unit at 40% blower speed and the power consumption of the ten active tiles. The power consumption of the active tiles is measured using an electricity wattmeter–watts up 57777 [132].

The power consumption of CRAC as a function of the blower speed is shown in Fig. B10 (a). The blue solid curve in Fig. B10 (b) gives the power consumption of ten active tiles as a function of active tile fan speed, while the green dashed curve represents the total power consumption in the second case and is obtained by adding the CRAC power consumption at 40% to the blue curve. From the figure it can be seen that the power consumption in the second case (combination of varying active tiles fan speed and CRAC blower speed fixed to 40%) is much lower than the first case (combination of passive tiles and varying CRAC blower speed).



(a)

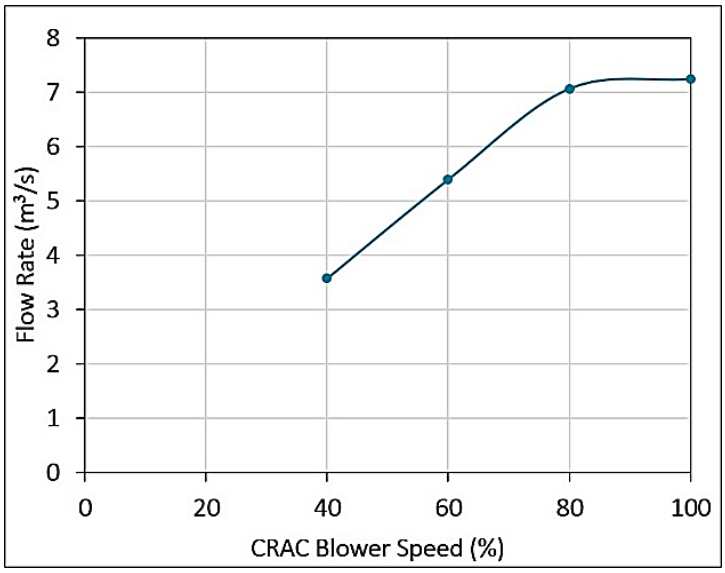


(b)

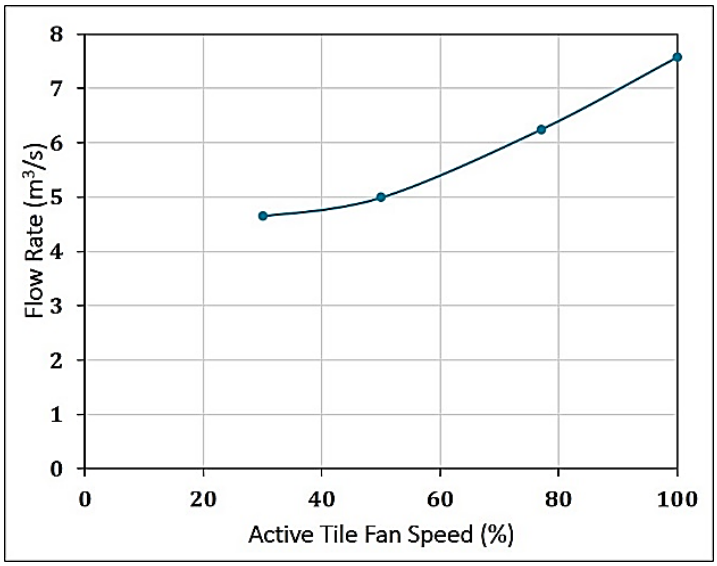
Figure B10 - (a) CRAC Power Consumption vs. CRAC Blower Speed (b) Total Power Consumption including 10 tiles and CRAC Power at 40% Blower Speed

The total flow rate through the cold aisle in the two cases is presented in Fig. B11. Figure B11 (a) gives the total flow rate through the cold aisle as a function of the CRAC blower speed while Fig. B11 (b) gives the total flow rate through the cold aisle as a function

of the active tile fan speed when the CRAC blower speed is held constant at 40%. Putting these results in perspective, from Figs. B10 and B11 we can see that the highest flow rate that can be achieved by using CRAC only is 7.23m^3/s, with an associated power consumption of 4550 W. However, the same flow rate can be achieved using a combination of active tiles and CRAC (run at 40%) with power consumption of only 1400 W—a reduction of more than two-thirds.



(a)



(b)

Figure B11 - (a) Total Aisle Flow Rate for Passive Tiles vs. CRAC blower Speed (b) Total Aisle Flow Rate for Active Tiles vs. Active Tile Fan Speed

The specific power consumption is usually defined as the power required to deliver one cubic meter per second of cold air:

$$Specific\ Power\ Consumption = \frac{Total\ Power\ Consumption}{Total\ Tile\ Air\ Flow\ Rate} \quad \text{(B.6)}$$

Figure B12 plots the specific power consumption for these two cases as a function of CRAC blower, or active tile fan, speeds, respectively. It can be observed from the figure that except for the one-point corresponding to 40% CRAC speed for case 1, the specific power consumption is always greater for case 1 (all passive tiles) as compared to case 2 (all active tiles). An aisle of active tiles causes the plenum pressure to be much lower than in the case of passive tiles. This greatly reduces the amount of cold air leakage due to a high plenum pressure. Thus, in equation B.6 the denominator increases much more rapidly than the numerator, reducing the specific power consumption in case of active tiles.

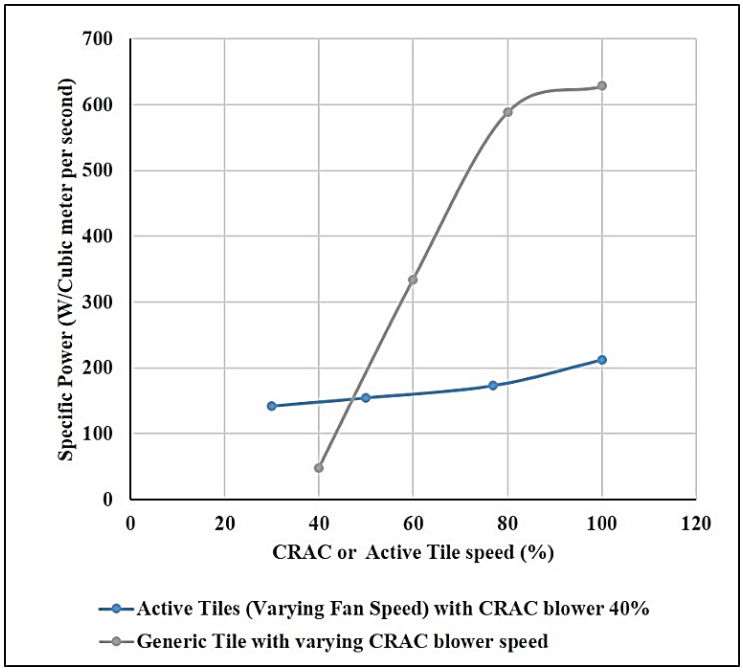


Figure B12 - Specific Power Consumption for aisle of Passive and Active Tiles

B.3.3 Cross-Correlation Experiments:

Active tiles can be used to mitigate isolated hot spots in the data center by supplying additional cold air when and where required. However, it is important to determine how varying the fan speed of a single active tile affects the adjacent tiles when implementing such a localized cooling strategy. Experiments were therefore also conducted to understand the qualitative relation (cross-correlation factors) between tile air flow rate and rack front face temperature for a group of three racks next to and above three active tiles when the active tiles systematically undergo step changes in their fan speeds. The experimental region of interest for this study, namely the three racks and three active tiles is highlighted in Fig. B13. The other seven tiles in the cold aisle are passive tiles. Each rack has a heat load of 10 kW, and the CRAC supply temperature and blower speed are set to be 15℃ and 80% respectively.

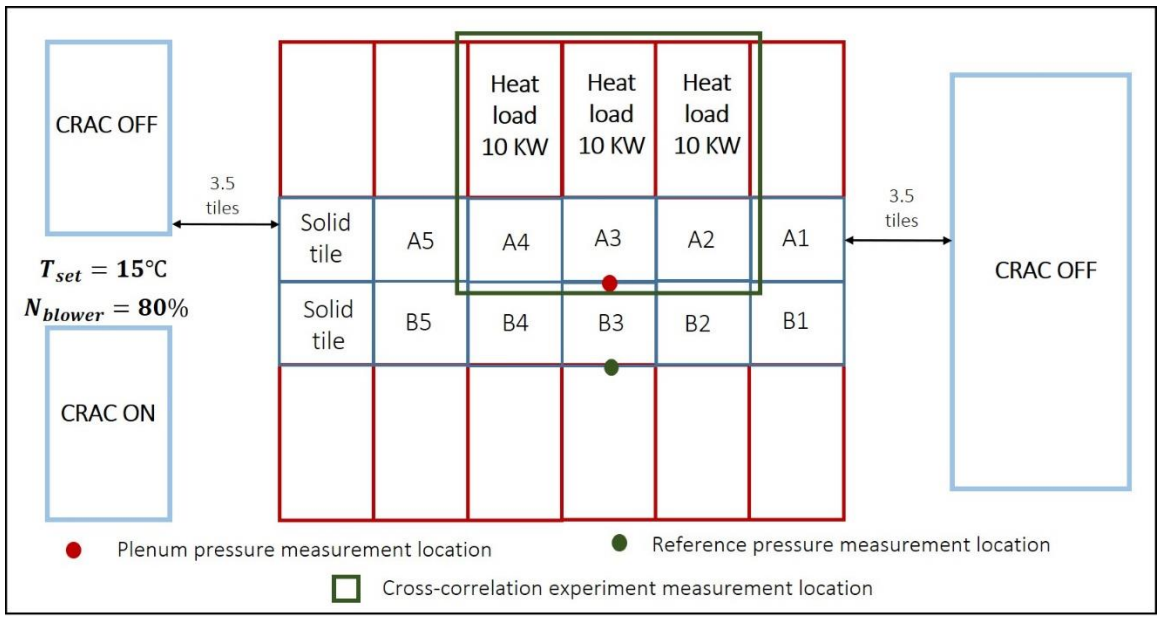


Figure B13- Experimental Section for Cross- Correlation Experiments

B.3.3.1 Tile Flow Rate Cross-Correlation

Figure B14 shows the air flow rate for the three tiles, as the fan speed for tile 2 is changed from 100% to 50% and then completely switched off. The fan speed is changed every 30 minutes to allow the conditions to achieve steady-state between these step changes in the fan speed. As the tile fan speed is decreased, the flow rate through tile 2 decreases rapidly. At the same time, the flow rate through the adjacent tiles also decreases slightly. This is because a step decrease in the active tile fan speed results in an increase in the relative plenum pressure (as the active tiles are extracting less cold air from the plenum).

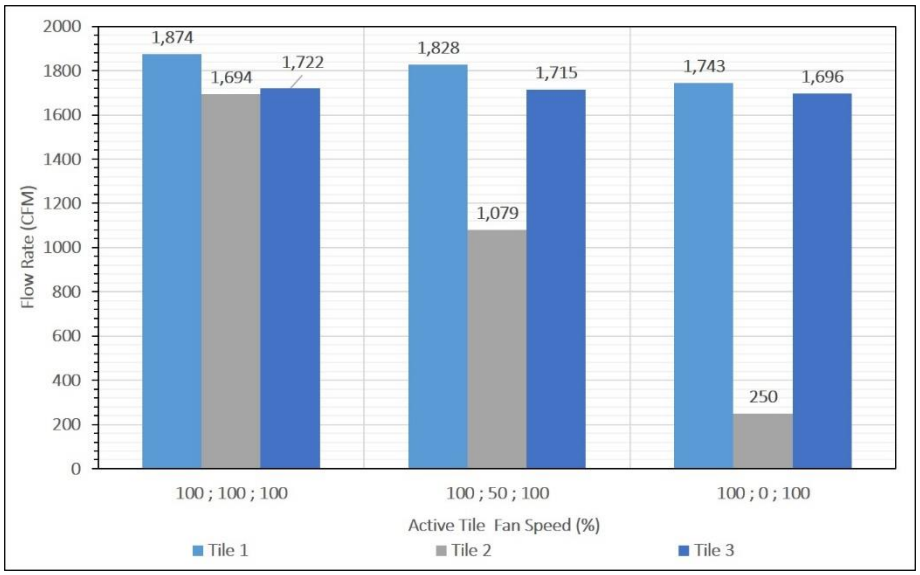


Figure B14 - Tile Flow Rate with Step Change in Active Tile Fan Speed

B.3.3.2 Temperature Cross-Correlation

In this study the fans for the three tiles were shut off in succession and the rack front face temperatures were monitored to gauge the effect of active tile fan speed on rack inlet temperature. Six T-type thermocouples were installed at the centerline of the rack front face along the height for each rack, so that the temperatures could be constantly and

simultaneously measured for all the three racks. A RFM SN802GRC-4M [133] transceiver module was used to wirelessly monitor temperatures. In the experiment, which lasted 120 minutes, the active tile fans first ran at 100% for the first 30 min. Then tile A4 was turned off for 30 minutes, followed by tile A3, thenA2. Figure B15 shows the temperature changes of the three racks upon step changes of the three tiles below. The three temperatures readings correspond to sensors from the three racks located at the same height of 35cm from the top of the rack. Note that all six sensors for each rack were monitored but only one is reported per rack. The time series in Fig. B15 shows that there is a negative correlation between the active tile fan speed (switched on) and the rack inlet temperature for each tile-rack pair, i.e., the inlet temperature decreases when the active tile is switched on. For the first 30 minutes the temperatures at the three sensors are almost constant and their mean values nearly identical. At t=30 min when the fan for tile 3 is switched off, the temperature for rack 3 spikes up quickly and stays at this high value until the fans for tile 3 are switched on again, after which the temperature quickly decreases to its earlier value. Qualitatively similar behaviors are observed when the tile 2 and tile 1 fans are switched off. Interestingly, shutting off the fans for tile 2 causes the temperatures of the racks 1 and 3 to increase. This behavior is not observed, however, for the other pairs. This may be related to the previous result wherein the flow rate through adjacent tiles decreased when the fan speed of the middle tile was decreased.

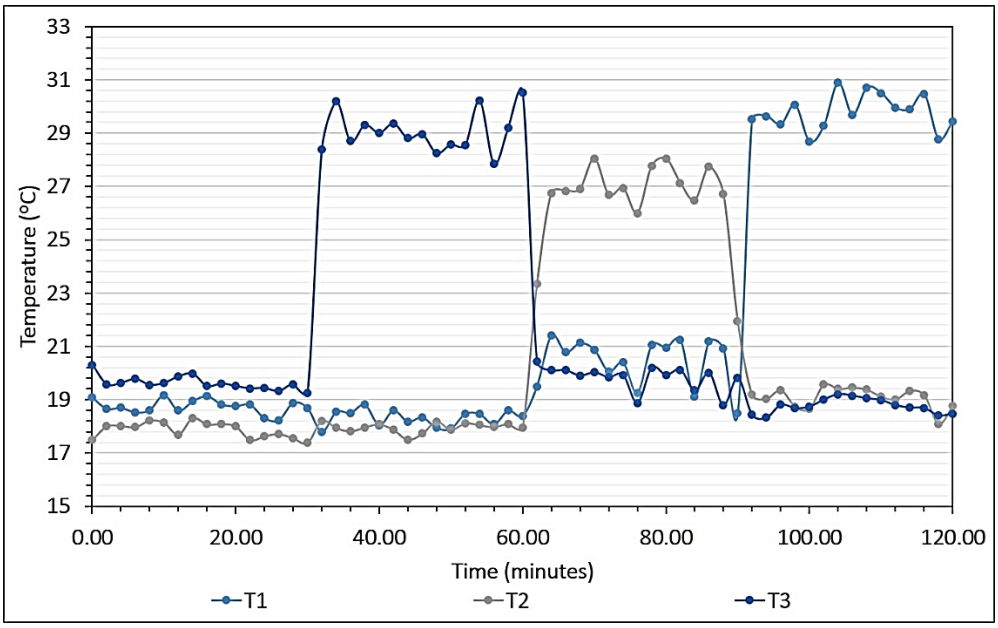


Figure B15 - Rack Front Face Temperature with Step Change in Corresponding Active Tile Fan Speed

Figure B15 demonstrates that the change in front face temperature for a rack when the fans of corresponding active tile is switched on is between 7 ℃ and 10 ℃ for the three tiles, occurring rapidly. This result suggests that active tiles are ideal candidates for implementing local cooling control in the data center where we need to rapidly mitigate hot spots at a particular location in an energy efficient manner. The following section proposes a potential control framework using active tiles.

B.4 Summary

This study examines the impact of active tiles on thermal and flow characteristics in a data center for two different configurations – single active tile and an aisle of active tiles. The results indicate that active tiles, as the actuators closest to the racks, can significantly and rapidly impact the local distribution of cooling resources. This achieves fine control of the tile supply flow rate and the rack inlet temperature and potentially lower energy consumption for the blowers. They could therefore be used in an appropriate control

framework to rapidly mitigate hot spots and maintain local conditions in an energy-efficient manner. Also, specific power consumption is lower when operating with an aisle of active tiles as against passive tiles due to lowered air leakage from the plenum space.

References

[1] "Http://Www.Datacenterdynamics.Com/Content-Tracks/Colo-Cloud/Number-of-Data-Centers-to-Decrease-after-2017/91495.Fullarticle," (2014).

[2] T. ASHRAE, "9.9 (2011) Thermal Guidelines for Data Processing Environments–Expanded Data Center Classes and Usage Guidance," *Whitepaper prepared by ASHRAE technical committee (TC),* vol. 9, (2011).

[3] "Data Center Thermal Runaway. A Review of Cooling Challenges in High Density Mission Critical Environments," in *White Paper by Active Power*.(2007)

[4] M. Stansberry and J. Kudritzki, "Uptime Institute 2012 Data Center Industry Survey," *Uptime Institute Survey,* (2012).

[5] "2013 Cost of Data Center Outages," in *Research Report by Ponemon Institute, Sponsored by Emerson Network Power*.(2013)

[6] M. Iyengar and R. Schmidt, "Analytical Modeling for Thermodynamic Characterization of Data Center Cooling Systems," *ASME Journal of Electronic Packaging,* vol. 131, p. 021009, (2009).

[7] C. Belady, D. Azevedo, M. Patterson, J. Pouehe, and R. Tipley, "Carbon Usage Effectiveness (Cue): A Green Grid Data Center Sustainability Metric. ," in *White Paper by The Green Grid*.(2010)

[8] M. Patterson, B. Tschudi, O. Vangeet, J. Cooley, and D. Azevedo, "Ere: A Metric for Measuring the Benefit of Reuse Energy from a Data Center," in *White Paper by The Green Grid*.(2010)

[9] D. Azevedo and S. C. Belady, "Water Usage Effectiveness (Wue™): A Green Grid Datacenter Sustainability Metric," in *White Paper by Green Grid*.(2011)

[10] A. Shehabi, S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, *et al.*, "United States Data Center Energy Usage Report," Lawrence Berkeley National Laboratory (LBNL)(2016).

[11] R. Ayoub, K. R. Indukuri, and T. S. Rosing, "Energy Efficient Proactive Thermal Management in Memory Subsystem," in *ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED)*, pp. 195-200.(2010)

[12] A. P. Ferreira, D. Mosse, and J. C. Oh, "Thermal Faults Modeling Using a Rc Model with an Application to Web Farms," in *19th Euromicro Conference on Real-Time Systems (ECRTS'07)*, pp. 113-124.(2007)

[13] X. S. Zhang and J. W. VanGilder, "Real-Time Data Center Transient Analysis," in *ASME Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS and NEMS: Volume 2, InterPACK*, pp. 471-477.(2011)

[14] S. L. Masson, D. Nörtershäuser, B. Deddy, and P. Glouannec, "Thermal Model for Data Centre Cooling," in *IEEE 33rd International Telecommunications Energy Conference (INTELEC)*, pp. 1-6.(2011)

[15] S. McAllister, V. P. Carey, A. Shah, C. Bash, and C. Patel, "Strategies for Effective Use of Exergy-Based Modeling of Data Center Thermal Management Systems," *Microelectronics Journal*, vol. 39, pp. 1023-1029, 2008/07/01/ (2008).

[16] X. Zhang, J. W. VanGilder, C. M. Healey, and Z. R. Sheffer, "Compact Modeling of Data Center Air Containment Systems," in *ASME. International Electronic Packaging Technical Conference and Exhibition, Volume 2: Thermal Management; Data Centers and Energy Efficient Electronic Systems*, p. V002T09A026.(2013)

[17] S. Kang, R. Schmidt, K. M. Kelkar, A. Radmehr, and S. V. Patankar, "A Methodology for the Design of Perforated Tiles in Raised Floor Data Centers Using Computational Flow Analysis," in *ITHERM, The Seventh Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 1-224.(2000)

[18] R. R. Schmidt, K. C. Karki, K. M. Kelkar, A. Radmehr, and S. V. Patankar, "Measurements and Predictions of the Flow Distribution through Perforated Tiles in Raised Floor Data Centers," *Proceedings of The Pacific Rim/ASME International Electronic Packaging Technical Conference and Exhibition, InterPACK,* (2001).

[19] R. Schmidt and E. Cruz, "Raised Floor Computer Data Center: Effect on Rack Inlet Temperatures of Chilled Air Exiting Both the Hot and Cold Aisles," in *Eighth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, ITherm* pp. 580-594.(2002)

[20] R. R. Schmidt and M. Iyengar, "Comparison between Underfloor Supply and Overhead Supply Ventilation Designs for Data Center High-Density Clusters," *ASHRAE Transactions*, vol. 113, pp. 115-125, (2007).

[21] K. C. Karki, A. Radmehr, and S. V. Patankar, "Use of Computational Fluid Dynamics for Calculating Flow Rates through Perforated Tiles in Raised-Floor Data Centers," *HVAC&R Research*, vol. 9, pp. 153-166, 2003/04/01 (2003).

[22] J. D. Rambo and Y. K. Joshi, "Multi-Scale Modeling of High Power Density Data Centers," in *International Electronic Packaging Technical Conference and Exhibition, Volume 1, InterPACK*, pp. 521-527.(2003)

[23] S. Bhopte, M. K. Iyengar, B. Sammakia, R. Schmidt, and D. Agonafer, "Numerical Modeling of Data Center Clusters: Impact of Model Complexity," *ASME International Mechanical Engineering Congress and Exposition, Heat Transfer, Volume 3* pp. 51-60, (2006).

[24] S. Bhopte, B. Sammakia, R. Schmidt, M. K. Iyengar, and D. Agonafer, "Effect of under Floor Blockages on Data Center Performance," in *Thermal and Thermomechanical Proceedings 10th Intersociety Conference on Phenomena in Electronics Systems ITHERM.*, pp. 426-433.(2006)

[25] E. Cruz, Y. Joshi, M. Iyengar, and R. Schmidt, "Comparison of Numerical Modeling to Experimental Data in a Small, Low Power Data Center Test Cell," presented at the ASME International Mechanical Engineering Congress and Exposition, (2009).

[26] M. Iyengar, R. R. Schmidt, H. Hamann, and J. VanGilder, "Comparison between Numerical and Experimental Temperature Distributions in a Small Data Center Test Cell," presented at the ASME. International Electronic Packaging Technical Conference and Exhibition, InterPACK Conference, Volume 1, (2007).

[27] W. A. Abdelmaksoud, H. E. Khalifa, T. Q. Dang, R. R. Schmidt, and M. Iyengar, "Improved Cfd Modeling of a Small Data Center Test Cell," in *12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 1-9.(2010)

[28] A. Radmehr, K. C. Karki, and S. V. Patankar, "Analysis of Airflow Distribution across a Front-to-Rear Server Rack," in *ASME. International Electronic Packaging Technical Conference and Exhibition, InterPACK Conference, Volume 1*, pp. 837-843.(2007)

[29] H. S. Erden, H. E. Khalifa, and R. R. Schmidt, "Room-Level Transient Cfd Modeling of Rack Shutdown," presented at the ASME. International Electronic Packaging Technical Conference and Exhibition, Volume 2: Thermal Management; Data Centers and Energy Efficient Electronic Systems, (2013).

[30] V. K. Arghode and Y. Joshi, "Rapid Modeling of Air Flow through Perforated Tiles in a Raised Floor Data Center," in *Fourteenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 1354-1365.(2014)

[31] J. W. VanGilder, X. Zhang, and S. K. Shrivastava, "Partially Decoupled Aisle Method for Estimating Rack-Cooling Performance in near-Real Time," in *ASME. International Electronic Packaging Technical Conference and Exhibition, InterPACK Conference, Volume 1* pp. 781-789.(2007)

[32] G. M. Nelson, "Development of an Experimentally-Validated Compact Model of a Server Rack," Georgia Institute of Technology , G. W. Woodruff School of Mechanical Engineering, (2007).

[33] J. Choi, Y. Kim, A. Sivasubramaniam, J. Srebric, Q. Wang, and J. Lee, "Modeling and Managing Thermal Profiles of Rack-Mounted Servers with Thermostat," in *IEEE 13th International Symposium on High Performance Computer Architecture*, pp. 205-215.(2007)

[34] Z. M. Pardey, "Proposal for Standard Compact Server Model for Transient Data Center Simulations," *ASHRAE Transactions*, vol. 121, p. 413, (2015).

[35] J. W. VanGilder and X. Zhang, "Coarse-Grid Cfd: The Effect of Grid Size on Data Center Modeling," *ASHRAE Transactions*, vol. 114, pp. 166-181, (2008).

[36] S. V. Patankar, "Airflow and Cooling in a Data Center," *ASME Journal of Heat Transfer*, vol. 132, pp. 073001-073001-17, (2010).

[37] E. Wibron, "Cfd Modeling of an Air-Cooled Data Center," CHALMERS University of Technology, Gothenburg/Sweden, (2015).

[38] S. Gondipalli, M. Ibrahim, S. Bhopte, B. Sammakia, B. Murray, K. Ghose, *et al.*, "Numerical Modeling of Data Center with Transient Boundary Conditions," in *12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 1-7.(2010)

[39] M. Ibrahim, S. Gondipalli, S. Bhopte, B. Sammakia, B. Murray, K. Ghose, *et al.*, "Numerical Modeling Approach to Dynamic Data Center Cooling," in *12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 1-7.(2010)

[40] A. M. Almoli, "Air Flow Management inside Data Centres," University of Leeds, Deaprtment of Mechanical Engineering, University of Leeds, (2013).

[41] E. Cruz and Y. Joshi, "Coupled Inviscid-Viscous Solution Method for Bounded Domains: Application to Data-Center Thermal Management," *International Journal of Heat and Mass Transfer*, vol. 85, pp. 181-194, 2015/06/01/ (2015).

[42] R. Ghosh, "Transient Reduced-Order Convective Heat Transfer Modeling for a Data Center," Georgia Institute of Technology - G. W. Woodruff School of Mechanical Engineering, (2013).

[43] J. Rambo and Y. Joshi, "Modeling of Data Center Airflow and Heat Transfer: State of the Art and Future Trends," *Distributed and Parallel Databases*, vol. 21, pp. 193-225, 2007/06/01 (2007).

[44] S. Alkharabsheh, J. Fernandes, B. Gebrehiwot, D. Agonafer, K. Ghose, A. Ortega, *et al.*, "A Brief Overview of Recent Developments in Thermal Management in Data Centers," *ASME Journal of Electronic Packaging*, vol. 137, pp. 040801-040801-19, (2015).

[45] Y. Fulpagare and A. Bhargav, "Advances in Data Center Thermal Management," *Renewable and Sustainable Energy Reviews,* vol. 43, pp. 981-996, 2015/03/01/ (2015).

[46] J. Athavale, M. Yoda, and Y. Joshi, "Thermal Modeling of Data Centers for Control and Energy Usage Optimization," in *Advances in Heat Transfer*, ed: Elsevier, p.^pp., (2018)

[47] K. Chen, D. M. Auslander, C. E. Bash, and C. D. Patel, "Local Temperature Control in Data Center Cooling: Part I, Correlation Matrix," *HP Enterprise Software and Systems Laboratory, Report No. HPL-2006-42,* (2006).

[48] K. Chen, C. E. Bash, D. M. Auslander, and C. D. Patel, "Local Temperature Control in Data Center Cooling: Part Ii, Statistical Analysis," *HP Enterprise Software and Systems Laboratory, Report No. HPL-2006-43,* (2006).

[49] M. H. Beitelmal, Z. Wang, C. Felix, C. Bash, C. Hoover, and A. McReynolds, "Local Cooling Control of Data Centers with Adaptive Vent Tiles," in *ASME. International Electronic Packaging Technical Conference and Exhibition, InterPACK Conference, Volume 2* pp. 645-652.(2009)

[50] Z. Wang, A. McReynolds, C. Felix, C. Bash, C. Hoover, M. Beitelmal, *et al.*, "Kratos: Automated Management of Cooling Capacity in Data Centers with Adaptive Vent Tiles," in *ASME International Mechanical Engineering Congress and Exposition, Volume 10: Mechanical Systems and Control, Parts A and B* pp. 269-278.(2009)

[51] Z. Wang, C. Bash, N. Tolia, M. Marwah, X. Zhu, and P. Ranganathan, "Optimal Fan Speed Control for Thermal Management of Servers," in *ASME 2009 InterPACK Conference collocated with the ASME 2009 Summer Heat Transfer Conference and the ASME 2009 3rd International Conference on Energy Sustainability*, pp. 709-719.(2009)

[52] V. Sundaralingam, Y. Joshi, and V. Arghode, "Controller to Regulate Maximum Server Cpu Temperatures in a Rack by Varying Crac Supply Air Temperatures," in *ASME International Mechanical Engineering Congress and Exposition, Volume 7: Fluids and Heat Transfer, Parts A, B, C, and D*, pp. 1703-1709.(2012)

[53] J. Athavale, Y. Joshi, M. Yoda, and W. Phelps, "Impact of Active Tiles on Data Center Flow and Temperature Distribution," in *15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 1162-1171.(2016)

[54] C. D. Patel, C. E. Bash, R. Sharma, A. Beitelmal, and C. G. Malone, "Smart Chip, System and Data Center Enabled by Advanced Flexible Cooling Resources," in *Semiconductor Thermal Measurement and Management IEEE Twenty First Annual IEEE Symposium, 2005.*, pp. 78-85.(2005)

[55] C. D. Patel, C. E. Bash, R. Sharma, M. Beitelmal, and R. Friedrich, "Smart Cooling of Data Centers," in *ASME. International Electronic Packaging Technical Conference and Exhibition*, pp. 129-137.(2003)

[56] C. Bash, C. D. Patel, and R. K. Sharma, "Dynamic Thermal Management of Air Cooled Data Centers," in *Thermal and Thermomechanical Phenomena in Electronics Systems, 2006. ITHERM'06. The Tenth Intersociety Conference on*, pp. 8 pp.-452.(2006)

[57] Z. Wang, C. Bash, C. Hoover, A. McReynolds, C. Felix, and R. Shih, "Integrated Management of Cooling Resources in Air-Cooled Data Centers," in *2010 IEEE International Conference on Automation Science and Engineering*, pp. 762-767.(2010)

[58] R. Zhou, Z. Wang, C. E. Bash, and A. McReynolds, "Modeling and Control for Cooling Management of Data Centers with Hot Aisle Containment," in *ASME International Mechanical Engineering Congress and Exposition, Volume 4: Energy Systems Analysis, Thermodynamics and Sustainability; Combustion Science and Engineering; Nanoengineering for Energy, Parts A and B*, pp. 739-746.(2011)

[59] N. Ahuja, C. W. Rego, S. Ahuja, Z. Shen, and S. Shrivastava, "Real Time Monitoring and Availability of Server Airflow for Efficient Data Center Cooling," in *29th IEEE Semiconductor Thermal Measurement and Management Symposium*, pp. 243-247.(2013)

[60] Z. Shu, Z. Tianyu, N. Ahuja, G. Refai-Ahmed, Z. Yongzhong, C. Guofeng, *et al.*, "Real Time Thermal Management Controller for Data Center," in *Fourteenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 1346-1353.(2014)

[61] T. D. Boucher, D. M. Auslander, C. E. Bash, C. C. Federspiel, and C. D. Patel, "Viability of Dynamic Cooling Control in a Data Center Environment," *ASME Journal of Electronic Packaging*, vol. 128, pp. 137-144, (2005).

[62] S. K. Shrivastava, J. W. VanGilder, and B. G. Sammakia, "Optimization of Cluster Cooling Performance for Data Centers," in *11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 1161-1166.(2008)

[63] H. E. Khalifa and D. W. Demetriou, "Energy Optimization of Air-Cooled Data Centers," *Journal of Thermal Science and Engineering Applications*, vol. 2, pp. 041005-041005-13, (2011).

[64] A. Shah, V. Carey, C. Bash, and C. Patel, "An Open-System, Exergy-Based Analysis of Major Data Center Thermal Management Components," in *IMAPS Advanced Technology Workshop ATW on Thermal Management THERM, Palo Alto, CA*.(2004)

[65] A. J. Shah, V. P. Carey, C. E. Bash, and C. D. Patel, "An Exergy-Based Control Strategy for Computer Room Air-Conditioning Units in Data Centers," in *ASME International Mechanical Engineering Congress and Exposition, Heat Transfer, Volume 1*, pp. 61-67.(2004)

[66] A. J. Shah, V. P. Carey, C. E. Bash, and C. D. Patel, "Exergy-Based Optimization Strategies for Multi-Component Data Center Thermal Management: Part I — Analysis," in *ASME International Electronic Packaging Technical Conference and Exhibition, Advances in Electronic Packaging, Parts A, B, and C* pp. 205-213.(2005)

[67] A. J. Shah, V. P. Carey, C. E. Bash, and C. D. Patel, "Exergy-Based Optimization Strategies for Multi-Component Data Center Thermal Management: Part Ii — Application and Validation," in *ASME International Electronic Packaging Technical Conference and Exhibition, Advances in Electronic Packaging, Parts A, B, and C*, pp. 215-224.(2005)

[68] A. J. Shah, V. P. Carey, C. E. Bash, and C. D. Patel, "Exergy Analysis of Data Center Thermal Management Systems," *ASME Journal of Heat Transfer*, vol. 130, pp. 021401-021401-10, (2008).

[69] G. Li, M. Li, S. Azarm, J. Rambo, and Y. Joshi, "Optimizing Thermal Design of Data Center Cabinets with a New Multi-Objective Genetic Algorithm," *Distributed and Parallel Databases*, vol. 21, pp. 167-192, 2007/06/01 (2007).

[70] S. K. Shrivastava, J. W. VanGilder, and B. G. Sammakia, "Data Center Cooling Prediction Using Artificial Neural Network," in *ASME. International Electronic Packaging Technical Conference and Exhibition, InterPACK Conference, Volume 1* pp. 765-771.(2007)

[71] E. Samadiani, Y. Joshi, J. K. Allen, and F. Mistree, "Adaptable Robust Design of Multi-Scale Convective Systems Applied to Energy Efficient Data Centers," *Numerical Heat Transfer, Part A: Applications*, vol. 57, pp. 69-100, 2010/01/29 (2010).

[72] *Http://Www.6sigmadcx.Com/*.

[73] D. King, M. Ross, M. Seymour, and T. Gregory, "Comparative Analysis of Data Center Design Showing the Benefits of Server Level Simulation Models," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2014 30th Annual*, pp. 193-196.(2014)

[74] X. Zhang, J. W. VanGilder, M. Iyengar, and R. R. Schmidt, "Effect of Rack Modeling Detail on the Numerical Results of a Data Center Test Cell," in *Thermal and Thermomechanical Phenomena in Electronic Systems, 2008. ITHERM 2008. 11th Intersociety Conference on*, pp. 1183-1190.(2008)

[75] V. Semin, M. Bardsley, O. Rosten, and C. Aldham, "Application of a Multilevel Unstructured Staggered Solver to Thermal Electronic Simulations," in *Thermal Measurement, Modeling & Management Symposium (SEMI-THERM), 2015 31st*, pp. 287-292.(2015)

[76] *Https://Www.Vertivco.Com/*.

[77] *Www.Degreec.Com/*.

[78] V. K. Arghode, T. Kang, Y. Joshi, W. Phelps, and M. Michaels, "Anemometric Tool for Air Flow Rate Measurement through Perforated Tiles in a Raised Floor Data Center," in *Thermal Measurement, Modeling & Management Symposium (SEMI-THERM), 2015 31st*, pp. 163-171.(2015)

[79] *Http://Www.Tsi.Com/Alnor-Micromanometer-Axd610/*.

[80] Z. Rongliang, W. Zhikui, A. McReynolds, C. E. Bash, T. W. Christian, and R. Shih, "Optimization and Control of Cooling Microgrids for Data Centers," in *13th InterSociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 338-343.(2012)

[81] D. P. Solomatine and A. Ostfeld, "Data-Driven Modelling: Some Past Experiences and New Approaches," *Journal of Hydroinformatics*, vol. 10, p. 3, (2008).

[82] J. D. Rambo, "Reduced-Order Modeling of Multiscale Turbulent Convection: Application to Data Center Thermal Management," Georgia Institute of Technology, G.W. Woodruff School of Mechanical Engineering, (2006).

[83] E. Samadiani and Y. Joshi, "Proper Orthogonal Decomposition for Reduced Order Thermal Modeling of Air Cooled Data Centers," *ASME Journal of Heat Transfer*, vol. 132, pp. 071402-071402-14, (2010).

[84] R. Ghosh and Y. Joshi, "Dynamic Reduced Order Thermal Modeling of Data Center Air Temperatures," in *ASME. International Electronic Packaging Technical Conference and Exhibition, ASME 2011 Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS and NEMS: Volume 2, InterPACK*, pp. 423-432.(2011)

[85] E. Samadiani, Y. Joshi, H. Hamann, M. K. Iyengar, S. Kamalsy, and J. Lacey, "Reduced Order Thermal Modeling of Data Centers Via Distributed Sensor Data," *ASME Journal of Heat Transfer*, vol. 134, pp. 041401-041401-8, (2012).

[86] Z. Song, B. T. Murray, and B. Sammakia, "Data Center Transient Flow Analysis Using Proper Orthogonal Decomposition," in *ASME International Mechanical Engineering Congress and Exposition, Volume 8B: Heat Transfer and Thermal Engineering*, p. V08BT09A040.(2013)

[87] R. Ghosh and Y. Joshi, "Rapid Temperature Predictions in Data Centers Using Multi-Parameter Proper Orthogonal Decomposition," *Numerical Heat Transfer, Part A: Applications,* vol. 66, pp. 41-63, 2014/07/01 (2014).

[88] Z. Song, B. T. Murray, and B. Sammakia, "Long-Term Transient Thermal Analysis Using Compact Models for Data Center Applications," *International Journal of Heat and Mass Transfer,* vol. 71, pp. 69-78, 2014/04/01/ (2014).

[89] J. Xu, M. Zhao, J. Fortes, R. Carpenter, and M. Yousif, "On the Use of Fuzzy Modeling in Virtualized Data Center Management," in *Fourth International Conference on Autonomic Computing (ICAC'07),* pp. 25-25.(2007)

[90] J. Moore, J. Chase, K. Farkas, and P. Ranganathan, "Data Center Workload Monitoring, Analysis, and Emulation," in *Eighth Workshop on Computer Architecture Evaluation using Commercial Workloads*, pp. 1-8.(2005)

[91] J. Moore, J. Chase, and P. Ranganathan, "Consil: Low-Cost Thermal Mapping of Data Centers," in *The First Workshop on Tackling Computer Systems Problems with Machine Learning (SysML)*.(2006)

[92] J. Moore, J. S. Chase, and P. Ranganathan, "Weatherman: Automated, Online and Predictive Thermal Mapping and Management for Data Centers," in *IEEE International Conference on Autonomic Computing*, pp. 155-164.(2006)

[93] Z. Song, B. T. Murray, and B. Sammakia, "Multivariate Prediction of Airflow and Temperature Distributions Using Artificial Neural Networks," in *ASME 2011 Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS and NEMS: Volume 2*, pp. 595-604.(2011)

[94] F. de Lorenzi and C. Vömel, "Neural Network-Based Prediction and Control of Air Flow in a Data Center," *ASME Journal of Thermal Science and Engineering Applications,* vol. 4, pp. 021005-021005-8, (2012).

[95] J. Gao and R. Jamidar, "Machine Learning Applications for Data Center Optimization," *Google White Paper,* (2014).

[96] L. da Fontoura Costa and G. Travieso, "Fundamentals of Neural Networks: By Laurene Fausett. Prentice-Hall, 1994, Pp. 461, Isbn 0-13-334186-0, Elsevier," ed: Elsevier,(1996)

[97] A. K. Jain, M. Jianchang, and K. M. Mohiuddin, "Artificial Neural Networks: A Tutorial," *Computer, Volume: 29, Issue: 3,* vol. 29, pp. 31-44, (1996).

[98] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-Propagating Errors," *nature,* vol. 323, p. 533, (1986).

[99] D. C. Montgomery, *Design and Analysis of Experiments*: John wiley & sons,(2017).

[100] M. Stein, "Large Sample Properties of Simulations Using Latin Hypercube Sampling," *Technometrics*, vol. 29, pp. 143-151, (1987).

[101] K. Mehrotra, C. K. Mohan, and S. Ranka, *Elements of Artificial Neural Networks*: MIT press,(1997).

[102] R. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE Assp magazine*, vol. 4, pp. 4-22, (1987).

[103] L. Zhang, J.-H. Jiang, P. Liu, Y.-Z. Liang, and R.-Q. Yu, "Multivariate Nonlinear Modelling of Fluorescence Data by Neural Network with Hidden Node Pruning Algorithm," *Analytica Chimica Acta*, vol. 344, pp. 29-39, (1997).

[104] G. Daqi and W. Shouyi, "An Optimization Method for the Topological Structures of Feed-Forward Multi-Layer Neural Networks," *Pattern recognition*, vol. 31, pp. 1337-1342, (1998).

[105] M. Dorofki, A. H. Elshafie, O. Jaafar, O. A. Karim, and S. Mastura, "Comparison of Artificial Neural Network Transfer Functions Abilities to Simulate Extreme Runoff Data," *International Proceedings of Chemical, Biological and Environmental Engineering*, vol. 33, pp. 39-44, (2012).

[106] H. Yu, Wilamowski, B.,, "Levenberg-Marquardt Training."

[107] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144-152.(1992)

[108] V. N. Vapnik, "An Overview of Statistical Learning Theory," *IEEE transactions on neural networks*, vol. 10, pp. 988-999, (1999).

[109] J. McMinis, "Intuition for Support Vector Regression and the Gaussian Process [Http://Mcminis1.Github.Io/Blog/2014/05/10/Intuition-for-Svr/]," (2014).

[110] J. McMinis, "Support Vector Regression and Gaussian Process Example [Http://Mcminis1.Github.Io/Blog/2014/05/26/Intuition-for-Svr-Ii/]," (2014).

[111] I. Guyon, "A Scaling Law for the Validation-Set Training-Set Size Ratio," *AT&T Bell Laboratories*, pp. 1-11, (1997).

[112] A. J. Smola and B. Schölkopf, "A Tutorial on Support Vector Regression," *Statistics and computing*, vol. 14, pp. 199-222, (2004).

[113] *Gaussian Process Regression Models [Https://Www.Mathworks.Com/Help/Stats/Gaussian-Process-Regression-Models.Html]*. 2017)

[114] M. Ebden, "Gaussian Processes for Regression : A Quick Introduction [Ftp://Ftp.Tuebingen.Mpg.De/Pub/Ebio/Chrisd/Gptutorial.Pdf]," (2008).

[115] C. E. Rasmussen, "Gaussian Processes in Machine Learning," in *Advanced Lectures on Machine Learning*, ed: Springer, p.^pp. 63-71. (2004)

[116] C. D. Patel, R. K. Sharma, C. E. Bash, and M. Beitelmal, "Energy Flow in the Information Technology Stack: Coefficient of Performance of the Ensemble and Its Impact on the Total Cost of Ownership," *Hewlett-Packard Laboratories, Palo Alto, CA, Technical Report No. HPL-2006-55*, (2006).

[117] T. J. Breen, E. J. Walsh, J. Punch, A. J. Shah, and C. E. Bash, "From Chip to Cooling Tower Data Center Modeling: Part I Influence of Server Inlet Temperature and Temperature Rise across Cabinet," in *12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 1-10.(2010)

[118] E. J. Walsh, T. J. Breen, J. Punch, A. J. Shah, and C. E. Bash, "From Chip to Cooling Tower Data Center Modeling: Part Ii Influence of Chip Temperature Control Philosophy," in *12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 1-7.(2010)

[119] V. K. Arghode, V. Sundaralingam, and Y. Joshi, "Airflow Management in a Contained Cold Aisle Using Active Fan Tiles for Energy Efficient Data-Center Operation," *Journal of Heat Transfer Engineering*, vol. 37, pp. 246-256, (2016).

[120] *Https://Www.Trane.Com/*.

[121] M. K. Patterson, "The Effect of Data Center Temperature on Energy Efficiency," in *11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 1167-1174.(2008)

[122] J. Athavale, Y. Joshi, and M. Yoda, "Artificial Neural Network Based Prediction of Temperature and Flow Profile in Data Centers," in *17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), San Diego, CA*.(2018)

[123] A. Konak, D. W. Coit, and A. E. Smith, "Multi-Objective Optimization Using Genetic Algorithms: A Tutorial," *Reliability Engineering & System Safety*, vol. 91, pp. 992-1007, 2006/09/01/ (2006).

[124] D. Kalyanmoy, "Multi-Objective Optimization Using Evolutionary Algorithms: An Introduction," *KanGAL Report*, (2011).

[125] T. Chow, G. Zhang, Z. Lin, and C. Song, "Global Optimization of Absorption Chiller System by Genetic Algorithm and Neural Network," *Energy and buildings*, vol. 34, pp. 103-109, (2002).

[126] L. Magnier and F. Haghighat, "Multiobjective Optimization of Building Design Using Trnsys Simulations, Genetic Algorithm, and Artificial Neural Network," *Building and Environment*, vol. 45, pp. 739-746, (2010).

[127] L. Zhou and F. Haghighat, "Optimization of Ventilation System Design and Operation in Office Environment, Part I: Methodology," *Building and Environment*, vol. 44, pp. 651-656, (2009).

[128] T. Weise, "Global Optimization Algorithms-Theory and Application," *Self-published*, vol. 2, (2009).

[129] N. El-Sayed, I. A. Stefanovici, G. Amvrosiadis, A. A. Hwang, and B. Schroeder, "Temperature Management in Data Centers: Why Some (Might) Like It Hot," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, pp. 163-174, (2012).

[130] I. Idelchik, *Flow Resistance: A Design Guide for Engineers*: Routledge,(2017).

[131] D. S. Miller, "Internal Flow Systems," (1978).

[132] *Http://Www.Powerwerx.Com/Techdata/Watts-up-V2.Pdf*.

[133] *Wireless.Murata.Com/*.